

Amal K. Mitra *Editor*

# Statistical Approaches for Epidemiology

From Concept to Application

# Statistical Approaches for Epidemiology

Amal K. Mitra  
Editor

# Statistical Approaches for Epidemiology

From Concept to Application



Springer

*Editor*

Amal K. Mitra 

Department of Epidemiology and Biostatistics  
Jackson State University  
Jackson, MS, USA

ISBN 978-3-031-41783-2

ISBN 978-3-031-41784-9 (eBook)

<https://doi.org/10.1007/978-3-031-41784-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2024

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Paper in this product is recyclable.

*I dedicate this book to my beloved family, whose unwavering love and support have been the foundation of my journey. To Ratna, my most beautiful and caring life partner, your constant encouragement has provided me with the strength to overcome obstacles and reach for the stars. To my two precious children, Amlan and Paromita, and my son-in-law, Tim, you are my greatest sources of inspiration. Your boundless curiosity, resilience, and unwavering belief in chasing dreams have motivated me to push my limits and strive for greatness. This dedication is a small token of my gratitude for the love, sacrifices, and endless support you have showered upon me. Thank you for standing by my side through thick and thin and for believing in me even when I doubted myself.*

*May this book serve as a tribute to the incredible individuals who have shaped me into the person I am today. Your love and presence have made all the difference, and I am forever grateful.*

# Preface

You will find many epidemiology textbooks available in the market. Why another? The reason is simple. This unique textbook provides readers with *basic concepts of epidemiology* while preparing them with *skills to apply statistical tools in real-life situations*. Students, in general, struggle with statistical theories and their practical applications. This book makes statistical concepts easy to understand by focusing on real-life examples, case studies, and exercises. The book also provides step-by-step guides for data analysis and interpretation using standard statistical software such as SPSS, SAS, R, Python, and GIS as appropriate, illustrating the concepts. In this book, you will find most of the things you need for a graduate-level course on epidemiology and research methods, all in one place.

What features distinguish this textbook from others?

- Practical examples of the use of the epidemiologic concepts
- Use of real-life data
- Plenty of step-by-step solutions to problems
- Use of statistical tools such as SPSS, SAS, Python, R, and GIS
- Special topics such as applications of artificial intelligence and machine learning, methods for handling missing data, systematic review and meta-analysis, tools for survival analysis, using SAS for multivariate analysis, clinical trials, steps of epidemic investigations, and many others
- An update on the epidemiology of COVID-19, and more importantly,
- Authorship with extensive teaching and research background

The 23 chapters of the book were chosen based on several factors: (1) the authors' and the editor's teaching and research experience in public health, the health sciences, and related disciplines; (2) the evolving nature of epidemiology, especially in the fields of data science, vaccine development, and epidemics due to infectious diseases, including COVID-19; and (3) a continued demand for a book of this nature from public health students and researchers from several countries with which the editor is affiliated.

Next, a little more perspective about the book's contents: Readers will primarily learn how to apply statistical methods in epidemiological studies and problem-solving. All the chapters first introduce the basic concepts related to the topic and then illustrate them using real-life examples. The book proceeds from an introduction to descriptive and analytical epidemiology (Chap. 1), to descriptions and applications of different types of epidemiologic studies (Chaps. 2, 3, and 4), descriptions of the epidemiologic measures of the distribution and determinants of diseases and events (Chap. 5), to clinical trials and ethical issues in human studies (Chap. 6), to the application of screening tools (Chap. 7), to surveillance methods (Chap. 8), to methods of standardization of rates (Chap. 9), and to the concept and application of Hill's criteria for causal association (Chap. 10). Readers will then learn about the concept of bias, confounding variables and effect modifiers, and methods of dealing with bias and confounders in research (Chap. 11), a step-by-step account of methods of investigation of an epidemic (Chap. 12), methods of population projection (Chap. 13), advanced techniques such as the use of GIS and spatial epidemiology (Chap. 14), survival analysis using SAS (Chap. 15), methods of systematic review and meta-analysis (Chap. 16), sample size estimation (Chap. 17), handling missing data (Chap. 18), and the use of prediction models such as artificial intelligence and machine learning using Python (Chap. 19). Chapters 20, 21, and 22 guide readers in using SPSS and SAS in data analysis illustrating real-life data. Finally, due to the COVID-19 pandemic, Chap. 23 describes the epidemiology of COVID-19 and offers techniques for assessing premature deaths due to the pandemic.

Each chapter is written by eminent scientists and experts worldwide, including contributors from institutions in the United States, Canada, Bangladesh, India, Hong Kong, Malaysia, and the Middle East. The book should be helpful in both developed and developing countries.

*Statistical Approaches for Epidemiology: From Concept to Application* is primarily targeted at graduate students, faculty, and researchers in public health and other branches of health sciences. However, the book is also helpful for graduate students and faculty in many related disciplines, such as data science, nursing, environmental health, occupational health, computer science, statistics, and biology.

My authors and I welcome your feedback, which will be valuable for future editions.

Jackson, MS, USA

Amal K. Mitra

# Contents

<b>1 Descriptive and Analytical Epidemiology .....</b>	1
Kiran Sapkota	
<b>2 Cross-Sectional Study: The Role of Observation in Epidemiological Studies .....</b>	19
Jean H. Kim	
<b>3 Case–Control Study .....</b>	43
Noraini Abdul Ghafar	
<b>4 Cohort Studies .....</b>	57
Deepa Valvi and Steven Browning	
<b>5 Epidemiological Measures .....</b>	77
Praphul Joshi	
<b>6 Clinical Trials .....</b>	91
Michael Bennish and Wasif Ali Khan	
<b>7 Screening .....</b>	115
Amal K. Mitra	
<b>8 Surveillance: The Role of Observation in Epidemiological Studies .....</b>	129
Adetoun F. Asala	
<b>9 Standardization .....</b>	147
Anwar T. Merchant	
<b>10 Causal Association .....</b>	155
Amal K. Mitra	
<b>11 Bias, Confounding, and Effect Modifier .....</b>	169
Dipak Kumar Mitra and Abdullah H. Baqui	

<b>12</b>	<b>Epidemic Investigation and Control . . . . .</b>	183
	Rajat Das Gupta and Sanjoy Kumar Sadhukhan	
<b>13</b>	<b>Population Projection . . . . .</b>	203
	Mohammad Mainul Islam and Amal K. Mitra	
<b>14</b>	<b>Geospatial Applications in Epidemiology: Location, Location . . . . .</b>	217
	Stephen Scroggins	
<b>15</b>	<b>Survival Analysis and Applications Using SAS and SPSS. . . . .</b>	235
	Rafiqul Chowdhury and Shahriar Huda	
<b>16</b>	<b>Systematic Review and Meta-Analysis: Evidence-Based Decision-Making in Public Health. . . . .</b>	257
	Aliyar Cyrus Fouladkhah and Minoo Bagheri	
<b>17</b>	<b>Sample Size Estimation . . . . .</b>	275
	Amal K. Mitra	
<b>18</b>	<b>Missing Data Imputation: A Practical Guide. . . . .</b>	293
	Enayetur Raheem	
<b>19</b>	<b>Artificial Intelligence and Machine Learning . . . . .</b>	317
	Hamidreza Moradi	
<b>20</b>	<b>A Step-by-Step Guide to Data Analysis Using SPSS: Iron Study Data . . . . .</b>	343
	Amal K. Mitra	
<b>21</b>	<b>Data Analysis Using SPSS: Jackson Heart Study . . . . .</b>	363
	Clifton C. Addison and Brenda W. Campbell Jenkins	
<b>22</b>	<b>Multiple Linear Regression and Logistic Regression Analysis Using SAS. . . . .</b>	381
	Azad R. Bhuiyan and Lei Zhang	
<b>23</b>	<b>Epidemiology of the COVID-19 Pandemic: An Update . . . . .</b>	411
	Amal K. Mitra	
<b>Index. . . . .</b>		427

# Contributors

**Clifton C. Addison** is an Associate Professor at Department of Biostatistics and Epidemiology, School of Public Health, Jackson State University (JSU). He is a Senior Research Scientist with the Jackson Heart Study (JHS) Graduate Training and Education Center (GTEC).

**Adetoun Faith Asala** is an Epidemiologist Team Lead at Mississippi State Department of Health, Office of Preventive Health, Mississippi. She serves as an adjunct professor at St. George's University School of Medicine, Department of Public Health and Preventive Medicine.

**Minoo Bagheri** is a research instructor in the Department of Cardiovascular Medicine at Vanderbilt University of Medical Sciences. She completed a predoc-toral and two postdoctoral fellowships at Harvard University, the University of Alabama in Birmingham, and at Vanderbilt University of Medical Sciences, respectively.

**Abdullah H. Baqui** is a Professor of Department of International Health and Director of the International Center for Maternal and Newborn Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD. He demonstrated the effectiveness of simple but effective strategies to reduce preventable newborn deaths in developing countries.

**Michael Bennish** is a pediatrician with subspecialty training in infectious diseases. He has been active in international health for 40 years and has conducted numerous clinical trials. He is currently the Executive Director of Mpilonhle, a nonprofit organization in rural South Africa.

**Azad R. Bhuiyan** is a Full Professor of the JSU School of Public Health. He is conducting research on health disparities and minority health in Mississippi. Earlier in his career, he served as a Research Associate of the Bogalusa Heart Study.

**Steven Browning** is an Associate Professor in the Department of Epidemiology in the College of Public Health of the University of Kentucky. Dr. Browning is also the Assistant Director of the Central Appalachian Regional Education and Research Center and an associate faculty member in the Kentucky Injury Prevention and Research Center.

**Rafiqul Chowdhury** is an Instructor at Department of Mathematics and Statistics, University of Fraser Valley Abbotsford, British Columbia, Canada. His major research includes multistage models for big data from repeated measures and prediction.

**Aliyar Cyrus Fouladkhah** is currently a tenured associate professor of Public Health Microbiology at Tennessee State University and the founding director of the Public Health Microbiology Foundation in Nashville. He is a graduate of Yale School of Public Health with a degree in Applied Biostatistics and Epidemiology.

**Noraini Abdul Ghafar** is a senior lecturer at the School of Health Sciences, Universiti Sains Malaysia. She graduated from the National University of Malaysia with a Ph.D. in Epidemiology and Biostatistics. She is currently teaching undergraduates and postgraduates courses, including Epidemiology, Biostatistics, and Research Methodology.

**Rajat Das Gupta** is medical graduate from Dhaka Medical College, University of Dhaka, and completed his MPH from Brac University, Dhaka, Bangladesh. He is a doctoral student in Epidemiology at Arnold School of Public Health, University of South Carolina.

**Shahriar Huda** received his Ph.D. in Statistics from Imperial College, London, United Kingdom. Currently, he is a Professor of Statistics at Kuwait University, Kuwait. Earlier, he served in several capacities in King Saud University, Saudi Arabia; University of Dhaka; and the Indian Statistical Institute in Kolkata.

**Mohammad Mainul Islam** received his Ph.D. in Demography from the Institute of Population Research, Peking University, China. He is currently a professor and former Chairman of the Department of Population Sciences at the University of Dhaka, Bangladesh. In 2010, he was named the Nick Simons Scholar of the New Investigator in Global Health Program, Washington, DC.

**Brenda W. Campbell Jenkins** is an investigator and Director of Training and Education with the Jackson Heart Study (JHS) Graduate Training and Education Center (GTEC). She has also been Principal Investigator of Project Health, a cardiovascular disease prevention and intervention program to address risk factors for cardiovascular diseases.

**Praphul Joshi** is a full professor and coordinator of graduate programs in the Department of Public Health at Sam Houston State University. Over the last 3 years, he has led the surveillance efforts for COVID-19 in Southeast Texas.

**Wasif Ali Khan** completed his MBBS from Chittagong Medical College, Bangladesh, and MHS from Johns Hopkins Bloomberg School of Public Health. He has been working as a Scientist at a premier health research institute in Bangladesh (icddr,b). His current research in collaboration with McGill University is on early detection of hepatocellular carcinoma and breast cancer.

**Jean H. Kim** has a background in both laboratory science and social sciences with training in molecular biology and anthropology. With a Ph.D. from Harvard University and with the Harvard University's Sinclair-Kennedy scholarship, she joined the Chinese University of Hong Kong and serves there as an Associate Professor of public health.

**Anwar T. Merchant** is a Professor of Epidemiology at the Department of Epidemiology and Biostatistics, University of South Carolina in Columbia, South Carolina. He completed a Doctor of Science with majors in Epidemiology and Nutrition, an MPH from Harvard University, and a Doctor of Dental Medicine from Shiraz University, Iran.

**Dipak Kumar Mitra** is a Professor of Epidemiology in the Department of Public Health of North South University, Bangladesh. He obtained his PhD from Johns Hopkins Bloomberg School of Public Health and an MPH from Harvard School of Public Health, USA. His research interests are to design and evaluate maternal, neonatal and child health, and nutrition interventions in developing countries.

**Hamidreza Moradi** is an assistant professor of data sciences at the University of Mississippi Medical Center. He received his Ph.D. degree in Computer Science in 2020 from the University of Texas at San Antonio. As a multidisciplinary researcher, his research interests include applied deep learning and artificial intelligence, model explainability, and cloud computing.

**Enayetur Raheem** holds a Ph.D. in Statistics from the University of Windsor, Ontario, Canada, and is currently serving as Head of Research, Biomedical Research Foundation, in Dhaka, Bangladesh. His research focuses on applying statistical and machine learning techniques in pharmaceutical manufacturing and drug discovery.

**Sanjoy Kumar Sadhukhan** is the Director and Professor in the Department of Epidemiology at All India Institute of Hygiene and Public Health (AIIPH&PH). He has 26 years of teaching and research experience in Public Health in India.

**Kiran Sapkota** after obtaining his Ph.D. in cancer epidemiology, joined the College of Health Sciences at Sam Houston State University in Huntsville, Texas, as an Assistant Professor in the Department of Public Health. His research focus is on chronic disease epidemiology, risk factors identification, immigrant health, and global health.

**Stephen Scroggins** is a spatial epidemiologist and researcher at Saint Louis University, College for Public Health and Social Justice, where his research focuses on exploring social and environmental factors impacting health outcomes. He received his Ph.D. in Public Health from Saint Louis University and M.Sc. degree in Health Sciences from Indiana State University.

**Deepa Valvi** is a Data Research Analyst in the Department of Surgery in the Division of Transplant Surgery at the University of Kentucky. She continues to coordinate clinical research. She has been an Asthma Epidemiologist at the State Department of Health, Frankfort, Kentucky.

**Lei Zhang** is Professor and Associate Dean for research and scholarship in the School of Nursing at the University of Mississippi Medical Center. Earlier, he served as director and chief research biostatistician of the Office of Health Data and Research at the Mississippi State Department of Health.

## About the Editor

**Amal K. Mitra, MD, MPH, DIH, DrPH,** is an internationally recognized scientist and leader in the field of public health. He has a dual background in medicine and in public health. He was a pioneer in developing a Master's degree program in Epidemiology and Biostatistics at The University of Southern Mississippi in 1998. He was the Founding Program Director of Public Health in the Faculty of Medicine at Kuwait University. Currently, Dr. Mitra is a tenured Professor of Epidemiology and Biostatistics at Jackson State University (JSU) in Jackson, Mississippi. He has been serving as the Director of Global Health Program Initiatives at JSU School of Public Health. He is collaborating with several institutions in Bangladesh, Hong Kong, India, Kuwait, Nepal, and Malaysia. Dr. Mitra is the recipient of many awards, including the Fulbright-Nehru Academic and Professional Excellence Award 2022–2023, Lifetime Achievement Award 2013, the Fulbright Scholar Award 2007–2008, the Innovation Award for Applied Research 2004, the Distinguished Teaching Award 2000, and the Distinguished Faculty Researcher Award 1999. He is the author of *Epidemiology for Dummies* published by Wiley & Sons in 2023.

# Abbreviations

AGRICOLA	AGRICultural Online Access
AGRIS	International System for Agricultural Science and Technology
AHI	Apnea-hypopnea index
AI	Artificial intelligence
AIDS	Acquired immunodeficiency syndrome
ANN	Artificial neural networks
ANOVA	Analysis of variance
aOR	Adjusted odds ratio
ARDS	Acute respiratory distress syndrome
ARIC	Atherosclerosis risk in communities
AUC	Area under the ROC curve
BIH	Bogalusa Heart Study
BIRPERHT	Bangladesh Institute for Research for Promotion of Essential and Reproductive Health Technologies
BMI	Body mass index
BPA	Bisphenol A
BRFSS	Behavioral risk factor surveillance system
BSE	Breast self-exam
CACE	Complier average causal effect
CCP	Cohort component method of projection
CDC	Centers for Disease Control and Prevention
CDR	Crude death rate
CFR	Case fatality rate
CHD	Coronary heart disease
CHS	Cardiovascular Health Study
CI	Confidence interval
CIOMS	Council for International Organizations of Medical Sciences
CNN	Convolutional neural networks
CONSORT	Consolidated Standards of Reporting Trials
COPD	Chronic obstructive pulmonary disease
COTC	Community-operated treatment center

COVID-19	Coronavirus disease 2019
CT	Computerized tomography
CV	Coefficient of variation
DL	Deep learning
DMSB	Data monitoring and safety board
DPCP	Detectable preclinical phase
EBMT	European Group for Blood and Marrow Transplantation
ECMO	Extracorporeal membrane oxygen
EDA	Exploratory data analysis
EM	Expectation-Maximization (EM) algorithm
FDA	Federal Drug Administration
FEV-1	Forced expiratory volume in 1 second
FIR	First information report
GBDT	Gradient Boosting Decision Tresses
GFR	Glomerular filtration rate
H2RA	Histamine 2 receptor antagonist
HBC	Hyperimmune bovine colostrum
HCC	Hepatocellular carcinoma
HFNC	Hospitalized and requires high-flow nasal- cannula oxygen
HIV	Human immunodeficiency virus
HPV	Human papillomavirus
HRT	Hormone replacement therapy
hs-CRP	high-sensitivity C-reactive protein
ICD	International Classification of Diseases
ICDDR,B	International Center for Diarrheal Disease Research, Bangladesh
ICH	International Conference (now Council) for Harmonization
IND	Investigational new drug
IRB	Institutional Review Board
JHS	Jackson Heart Study
KM analysis	Kaplan-Meier analysis
KNN	K-nearest neighbor
K-S test	Kolmogorov-Smirnov test
LDCT	Low-dose computed tomography
LISA	Local indicator of spatial autocorrelation
LOCF	Last observation carried forward
LSTM	Long short-term memory
MAR	Missing at random
MCAR	Missing completely at random
MERS	Middle East respiratory syndrome
ML	Machine learning
MLF	Maximum likelihood function
MMR	Mumps, measles, rubella
MNAR	Missing not at random
MRI	Magnetic resonance imaging
mRNA	Messenger ribonucleic acid

NCHS	National Center for Health Statistics
NHANES	National Health and Nutrition Examination Survey
NHIS	National Health Interview Survey
NHLBI	The National Heart, Lung and Blood Institute
NHS	Nurses' Health Study
NIH	National Institutes of Health
NIMHD	The National Institute on Minority Health and Health Disparities
NLP	Natural language processing
NN	Neural networks
NOCB	Next observation carried backward
NSAIDs	Non-steroidal anti-inflammatory drugs
OMB	Office of Management and Budget
ONS	The United Kingdom Office for National Statistics
OR	Odds ratio
OSA	Obstructive sleep apnea
PCR	Polymerase chain reaction
PH model	Proportional Hazards model
PKU	Phenylketonuria
PM 2.5	Particulate matter 2.5
PMR	Proportional mortality ratio
PPI	Proton pump inhibitor
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSA	Prostate-specific antigen
PYLL	Potential years of life lost
Q-Q plot	Quantile-quantile plot
RCT	Randomized controlled trial
REACT	Real-time assessment of community transmission
RF	Random forests
RNN	Recurrent neural networks
ROC	Receiver operating characteristic
RR	Relative risk
RRR	Relative risk ratio
RRT	Rapid response team
SARS	Severe acute respiratory syndrome
SARS-CoV-2	Severe acute respiratory syndrome coronavirus
SHAP	SHapley Additive exPlanation
SJMC	Subang Jaya Medical Centre
SMD	Standardized mean difference
SMR	Standardized mortality ratio
SMART	Simple, measurable, achievable, relevant, and time-bound
SOPs	Standard operating procedures
STI	Sexually transmitted infection
S-W test	Shapiro–Wilk test
TB	Tuberculosis

TMH	Tata Memorial Hospital
TPCP	Total preclinical phase
UMMC	Universiti Malaya Medical Centre; University of Mississippi Medical Center
USDA	The United States Department of Agriculture
USPSTF	The United States Preventive Services Task Force
UV	Ultraviolet
VIF	Variance inflation factor
WHO	World Health Organization
WIC	Special supplemental nutrition program for women, infants, and children
WUS	Wake up stroke
YPLL	Years of potential life lost
YRBSS	Youth Risk Behavior Surveillance System

# List of Figures

Fig. 1.1	A person–place–time model.....	2
Fig. 1.2	Types of epidemiological studies. Used with permission from John Wiley & Sons, Inc. from Chapter 17: Investigating the Types of Epidemiologic Studies, Mitra AK, first edition, 2023; permission conveyed through Copyright Clearance Center, Inc. [11] .....	8
Fig. 2.1	An overview of a cross-sectional study design.....	27
Fig. 2.2	Prevalence of obesity among US adults (in 1990, 2000, and 2010). (Source: <a href="https://www.cdc.gov/obesity/data/prevalence-maps.html#downloads">https://www.cdc.gov/obesity/data/prevalence-maps.html#downloads</a> ) .....	29
Fig. 3.1	Case–control study design [5]. Used with permission of John Wiley & Sons, Inc. from Chapter 17: Investigating the Types of Epidemiologic Studies, Mitra AK, first edition, 2023; permission conveyed through Copyright Clearance Center, Inc .....	44
Fig. 4.1	Design of a cohort study.....	59
Fig. 4.2	Prospective cohort study.....	61
Fig. 4.3	Retrospective cohort study .....	63
Fig. 5.1	Estimated teen birth rates for females aged 15–19 years by county in the United States for 2020. (Source: The National Center for Health Statistics, CDC [3]) .....	81
Fig. 5.2	Teen births by state in the United States for 2020. (Source: The National Center for Health Statistics, CDC [3]).....	82
Fig. 5.3	Number of births to unmarried women by age group in United States from 1940 to 2015. (Source: Ventura and Bachrach, 2000 [2]) .....	83
Fig. 5.4	Comparison of overall deaths, crude mortality rates, and age-adjusted mortality rates in the United States from 1935 to 2010. (Source: Centers for Disease Control and Prevention [3]) .....	84

Fig. 5.5	Infant mortality rates in the United States across the states for the year 2020. (Source: National Center for Health Statistics, CDC (2000) [3]).....	85
Fig. 5.6	Differences between crude and adjusted cause-specific mortality rates due to chronic obstructive pulmonary disease (COPD) in the United States from 2001 to 2020. (Data Source: CDC Wonder (wonder.cdc.gov) [4]) .....	85
Fig. 5.7	Trends in COPD mortality rates by 10-year age groups from 2001 to 2020. (Data source: CDC Wonder (wonder.cdc.gov) [4]).....	86
Fig. 5.8	Years of potential life lost (YPLL) due to Covid-19 by gender in Brazil in 2020.(Source: Castro et al. 2021 [5]) .....	87
Fig. 6.1	CONSORT flow diagram of the progress through the phases of a parallel randomized trial of two groups [30].....	104
Fig. 7.1	Validity and reliability [2]. (Used with permission of John Wiley & Sons, Inc. from Chapter 15: Identifying Diseases by Screening, Mitra AK, first edition, 2023; permission conveyed through Copyright Clearance Center, Inc.) .....	118
Fig. 7.2	A hypothetical model of cervical cancer .....	120
Fig. 7.3	Example of an ROC curve using hypothetical data.....	123
Fig. 7.4	Effect of changing the cutoff point on sensitivity and specificity of the test [2]. (Used with permission of John Wiley & Sons, Inc. from Chapter 15: Identifying Diseases by Screening, Mitra AK, first edition, 2023; permission conveyed through Copyright Clearance Center, Inc.) .....	126
Fig. 8.1	Newsworthy headlines highlighting past epidemics which public health surveillance provided information about. (Source – Centers for Disease Control and Prevention [1]) .....	131
Fig. 8.2	Processes involved in conducting a surveillance .....	133
Fig. 8.3	Public health approach to surveillance summarized. (Source – Centers for Disease Control and Prevention [1]) .....	135
Fig. 8.4	Disease reporting system for surveillance programs.....	137
Fig. 8.5	A public health surveillance system .....	141
Fig. 9.1	Age distribution of US population in 1978 [1].....	148
Fig. 9.2	Age distribution of US population in 2019 [2].....	149
Fig. 10.1	A schematic diagram showing the effect of a confounder .....	156
Fig. 10.2	A schematic diagram of Rothman's causal pie, modified [13].....	164
Fig. 11.1	Effect of confounding .....	174
Fig. 11.2	Effect of confounding eliminated .....	175

Fig. 11.3	Flow diagram to assess confounding and effect modification.....	180
Fig. 12.1	Epidemiological triad .....	186
Fig. 12.2	A schematic diagram of different steps of an outbreak investigation.....	189
Fig. 12.3	Epidemic curve of common source origin.....	192
Fig. 12.4	Epidemic curve of propagated source origin.....	193
Fig. 12.5	Number of patients treated by week, at the Shataki and Kalirbazar Community-Operated Treatment Centers (COTC), between 1 July and 31 December 1982 [10].....	199
Fig. 13.1	Example of UN projected world population, 1950–2100 [4].....	205
Fig. 13.2	Probabilistic projections of crude death rate (CDR) in the world. (Source: United Nations. Department of Economic and Social Affairs, Population Division. World Population Prospects, 2022 [4]) .....	208
Fig. 13.3	Probabilistic Projections on Life Expectancy of Both Sexes in the World. (Source: United Nations. Department of Economic and Social Affairs, Population Division. World Population Prospects, 2022 [4]) .....	209
Fig. 14.1	The socioecological model.....	218
Fig. 14.2	Example of autocorrelation among varying shades of blue .....	219
Fig. 14.3	Examples of queen and rook contiguity .....	224
Fig. 14.4	Depiction of spatial autocorrelation .....	226
Fig. 14.5	An example of a choropleth map. (Data source: U.S. Census Bureau [35]) .....	227
Fig. 14.6	An example of spatial data .....	228
Fig. 14.7	Choropleth map of the average restaurant in each neighborhood ....	229
Fig. 15.1	Relapse and death after the bone marrow transplantation .....	236
Fig. 15.2	Different types of censoring experienced by the six study subjects .....	237
Fig. 15.3	Two survival curves .....	240
Fig. 15.4	Kaplan–Meier estimate.....	244
Fig. 15.5	Kaplan–Meier estimate of the survival function curve.....	245
Fig. 15.6	The survival and hazard function curves by age at marriage .....	246
Fig. 15.7	Outputs from the Cox PH model using SAS PROC PHREG.....	249
Fig. 15.8	Log-rank test using SPSS .....	252
Fig. 16.1	PRISMA flow chart for illustration of searched article and inclusion process [18] (Reprinted with permission from the authors) .....	262
Fig. 16.2	Forest plot for the effect of physical activity label on calorie reduction by study setting [28] (Reprinted with permission from the authors) .....	263

Fig. 16.3	Non-linear dose-response plots on the relationship between body mass index (BMI) and mortality in stroke patients. The association of BMI with all-cause mortality (a). The association of BMI with stroke specific mortality (b). Continuous black and medium-dashed orange-red lines represent non-linear and linear plots. 95% confidence intervals are shown by long-dashed maroon lines. Vertical axes are based on the log-scale of the hazard ratios [29]. (Used with permission from the publisher; permission conveyed through Copyright Clearance Center, Inc.) .....	264
Fig. 16.4	Risk of bias graph: review authors' judgements about each risk of bias item presented as percentages across all included studies. (Reprinted with permission from the author) .....	268
Fig. 17.1	Power curve generated by G*Power.....	289
Fig. 18.1	Percentage of missing values by each variable .....	298
Fig. 18.2	Scatterplot of hemoglobin and ferritin .....	301
Fig. 18.3	Scatterplot of hemoglobin against ferritin by survival status .....	302
Fig. 18.4	Density plot of a continuous variable (hemoglobin) by survival status.....	303
Fig. 18.5	Distribution of iron after mean imputation.....	306
Fig. 18.6	Distribution of iron after regression imputation.....	307
Fig. 18.7	Distribution of iron after mean imputation followed by a regression imputation.....	308
Fig. 19.1	Regression line with actual and predicted observations .....	319
Fig. 19.2	A decision tree to predict patients' need for a prescription.....	323
Fig. 19.3	Three clusters of data in a two-dimensional space .....	325
Fig. 19.4	Clusters found by both K-means and OPTICS .....	327
Fig. 19.5	K-means clustering and synthetic data .....	328
Fig. 19.6	OPTICS clustering and synthetic data.....	329
Fig. 19.7	A simple deep neural network .....	330
Fig. 19.8	A single convolutional filter convolving over an image .....	336
Fig. 19.9	A max pooling filter convolving over an image .....	336
Fig. 20.1	Types of skewness [6].....	350
Fig. 20.2	Shapes of mesokurtic, leptokutic, and platykurtic curves [6] .....	350
Fig. 20.3	A histogram showing normal distribution of hemoglobin at baseline. <i>Note:</i> There is no skewness. Kurtosis is low [1].....	351
Fig. 20.4	Boxplot distribution of hemoglobin at baseline at three treatment groups .....	353
Fig. 20.5	A Stem-and-Leaf pattern of hemoglobin levels .....	354
Fig. 20.6	A Q–Q distribution showing normal (a), skewed distribution (b, c), thick (d) and thin (e) tailed plot.....	355
Fig. 20.7	A Q–Q distribution shows normal distribution of hemoglobin values [1] .....	356

Fig. 21.1	Steps showing the frequency distribution analysis. (Source: This figure is a screenshot of the SPSS data analysis using JHS data at our lab [2, 3]. This is generated when variables are entered as an initial step in analyzing the frequency distribution) .....	366
Fig. 21.2	Steps showing the descriptive data analysis. (Source: This figure is a screenshot of the SPSS data analysis using JHS data at our lab [2, 3]. This is generated when variables are entered as an initial step of analyzing descriptive statistics).....	366
Fig. 21.3	Steps showing the independent samples <i>t</i> -test. (Source: This figure is a screenshot of the SPSS data analysis for independent sample <i>t</i> -test with JHS data [2, 3]. This is generated when variable BMI is entered as a test variable and GENDER is used as a Grouping Variable for the test. This figure is a screen print of the SPSS page that is generated when variable (BMI) is entered for an independent sample <i>t</i> -test with GENDER) .....	367
Fig. 21.4	Steps showing the one-way analysis of variance (ANOVA) test. (Source: This figure is a screenshot of the SPSS data analysis for one-way ANOVA test using JHS data [2, 3]. Here, fasting glucose is a dependent variable and BMIGROUP is an independent factor used for the analysis).....	367
Fig. 21.5	Steps showing the bivariate Pearson correlation test. (Source: This figure is a screenshot of the SPSS data analysis for bivariate correlation test using JHS data [2, 3]. This is generated when variables such as BMI and fasting glucose are entered for the analysis) .....	368
Fig. 22.1	Quantile plot and histogram of the dependent variable.....	385
Fig. 22.2	Influence diagnostics .....	399
Fig. 22.3	Displayed ROC for model .....	401

# List of Tables

Table 1.1	Racial and ethnic categories in the United States.....	4
Table 1.2	Comparison of descriptive and analytical epidemiology .....	12
Table 1.3	A $2 \times 2$ contingency table for odds ratio .....	12
Table 1.4	Interpretation of odds ratio (OR) values in epidemiological studies.....	12
Table 2.1	Types of study designs and examples of research questions that may be answered .....	21
Table 2.2	Categories of risk factors and protective factors and methods of data collection .....	25
Table 2.3	Example of a 2-by-2 contingency table for calculating prevalence by exposure levels .....	30
Table 2.4	Calculating the odds ratio from contingency tables .....	32
Table 2.5	An example of misclassification in a cross-sectional study leading to erroneous conclusions .....	35
Table 3.1	Calculation of odds ratio for case-control study .....	51
Table 3.2	Food poisoning and contaminated salad.....	52
Table 4.1	Calculation of SMR for uranium workers .....	60
Table 4.2	Advantages and disadvantages of cohort studies .....	64
Table 4.3	Population cohorts of the Framingham Heart Study .....	65
Table 4.4	Interpretation of relative risk .....	71
Table 6.1	Phases of a clinical trial of a new drug, biological process, or device .....	93
Table 6.2	Ethical responsibilities of a principal investigator in a clinical trial.....	107
Table 6.3	Elements of informed consent form [37, 39].....	108
Table 7.1	Sensitivity, specificity, predictive values, and agreement of a screening test .....	119
Table 7.2	Data of screening test and the gold standard.....	124
Table 8.1	Differences between public health surveillance and epidemiologic research .....	136

Table 9.1 Deaths by age group in the USA in 2019 .....	149
Table 9.2 Direct standardization using US 1978 population as standard .....	150
Table 11.1 Different types of confounding effect: a hypothetical example .....	177
Table 11.2a Effect modification additive and ratio scales.....	178
Table 11.2b Effect modification additive and ratio scales.....	178
Table 11.2c Effect modification additive and ratio scales.....	179
Table 11.3 Quantitative effect modification .....	179
Table 11.4 Qualitative effect modification .....	180
Table 12.1 Trigger levels, their significance, and level of response of outbreak used in India [5] .....	187
Table 14.1 Levels of the socioecological model and the varying levels of locational influence .....	218
Table 14.2 The five geographical themes and respective epidemiological examples .....	220
Table 15.1 Data and layout for the six subjects.....	238
Table 15.2 Data for some selected subjects.....	243
Table 15.3 Distribution of hemorrhage by age at marriage.....	247
Table 15.4 Test of equality of survival distributions for the categories of age at marriage .....	247
Table 17.1 A $2 \times 2$ contingency table.....	278
Table 17.2 Common statistical parameters used for sample size estimation ....	281
Table 17.3 Use of the data for $f(\alpha, \beta)$ [5] .....	283
Table 17.4 $2 \times 2$ contingency table.....	287
Table 18.1 Variable description .....	295
Table 18.2 Bivariate analysis of risk factors and one-year survival status .....	296
Table 18.3 Testing MCAR for esophageal varices by one-year survival .....	299
Table 18.4 Testing for statistical difference of hemoglobin in the two outcome groups .....	302
Table 18.5 Bivariate analysis of risk factors and one-year survival status on the imputed data after KNN imputation .....	311
Table 22.1 Univariate analysis of dependent variable .....	384
Table 22.2 Partial display of influential statistics.....	386
Table 22.3 Collinearity diagnosis statistics .....	387
Table 22.4 ANOVA statistics.....	388
Table 22.5 Displayed the $R^2$ .....	389
Table 22.6 Displayed the parameter estimates of hs-CRP .....	389
Table 22.7a Forward procedure.....	392
Table 22.7b Backward procedure .....	393
Table 22.7c Stepwise procedure.....	394
Table 22.8 Summary of findings of the model selection.....	395
Table 22.9 Displayed the interaction term .....	397

Table 22.10 Displayed correlation matrix .....	398
Table 22.11 Displayed regression diagnostic .....	399
Table 22.12 Model fit statistics .....	400
Table 22.13 Improvement test .....	400
Table 22.14 Displayed Hosmer–Lemeshow statistics .....	402
Table 22.15 Displayed R-square statistics.....	403
Table 22.16 Maximum likelihood estimates .....	403
Table 22.17 Odds Ratio statistics .....	404
Table 22.18 Displayed trend analysis statistics .....	405
Table 22.19 Trend analysis for ordinal predictor .....	406
Table 23.1 Potential years of life lost (PYLL) due to COVID-19 in the United States (data as of December 28, 2022) .....	420
Table 23.2 Therapeutic management of nonhospitalized adults with COVID-19 .....	420
Table 23.3 Therapeutic management of adults hospitalized for COVID-19.....	421

# Chapter 1

## Descriptive and Analytical Epidemiology



Kiran Sapkota

### Learning Objectives

After completing this chapter, you will be able to:

- Describe the distribution of diseases by person, place, and time.
- Analyze and interpret several types of descriptive epidemiological studies.
- Describe distinct types of analytical studies.
- Evaluate the strength and limitations of different epidemiological studies.

## 1 Introduction

This chapter provides an overview of two types of epidemiological studies: descriptive and analytical. Further details of each of the study designs are illustrated in the next few chapters. Descriptive epidemiology is an essential aspect of epidemiological investigations, and it focuses on answering the “who, where, and when” questions of disease occurrence. Person, place, and time are the three primary characteristics that are used to understand the distribution and determinants of diseases in populations. Person-related factors in epidemiological investigations gather data on who is affected or has a disease. These factors include age, gender, race, marital status, socioeconomic status, occupation, and education. Epidemiologists can identify patterns and risk factors for different diseases by examining these factors. For example, certain diseases may be more common in specific age groups or among individuals with particular occupations. Place-related factors help understand where diseases or health conditions frequently occur.

---

K. Sapkota (✉)

Department of Public Health, College of Health Sciences, Sam Houston State University,  
Huntsville, TX, USA  
e-mail: [kxs133@shsu.edu](mailto:kxs133@shsu.edu)

Disease occurrence differs by geographic location, and the different aspects of location impact the distribution of disease. Diseases can occur locally, nationally, or even globally. Some diseases have different distribution patterns in urban or rural areas based on the variation of geography, environment, and demographics. Similarly, disease distribution patterns differ by time-related factors. The time for infectious diseases to occur or manifest can be short, ranging from a few hours to days. Some diseases, such as chronic diseases, take months to years to develop. Some infectious diseases occur at specific times, whereas others show cyclic patterns. In descriptive epidemiological studies, some health conditions can be studied from one time frame to another. By examining these characteristics, epidemiologists can design and implement effective prevention and control measures to reduce the disease burden in populations.

The second type of epidemiological study is called analytical study. Analytical epidemiological studies include ecological studies (correlational studies), analytical cross-sectional studies, analytical cohort studies, and experimental studies.

## 2 Person, Place, and Time Model

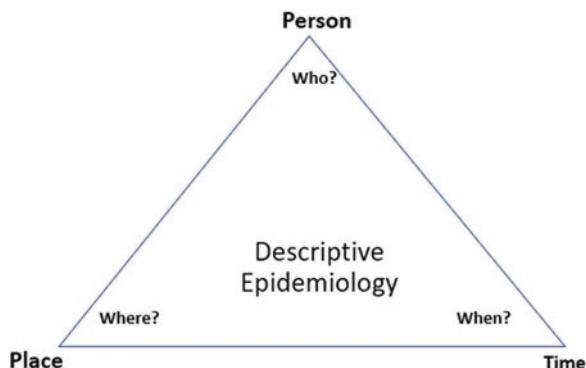
Figure 1.1 shows a descriptive model of person, place, and time.

### 2.1 Person-Related Factors

#### 2.1.1 Age

Age is a key factor for understanding the epidemiology of various diseases. The risk of developing diseases and the severity of those diseases often vary based on the individual's age. For example, some diseases occur primarily in children, whereas others are more common in adults or the elderly.

**Fig. 1.1** A person–place–time model



Infectious diseases such as measles, chickenpox, and influenza are common in children. Children's immune systems are still in the process of developing, making them more susceptible to infections. Childhood is also the time when many birth defects and genetic disorders are first identified. Additionally, unintentional injuries, such as falls and drowning, are the leading cause of death in this age group.

Teens are more prone to sexually transmitted infections (STIs), as they are at the age when they begin engaging in sexual activities. STIs such as chlamydia, gonorrhoea, and human papillomavirus (HPV) are common in this age group. Substance abuse is also more prevalent in teenagers, leading to various health issues, such as addiction and mental health problems.

Chronic diseases like hypertension, kidney disease, and chronic obstructive pulmonary disease (COPD) typically occur in middle or late adulthood. These diseases are often the result of lifestyle factors, such as poor diet, lack of exercise, and smoking. Similarly, some cancers, such as breast cancer, prostate cancer, and colon cancer, are more likely to develop later in life. Type 2 diabetes, which is associated with obesity and sedentary lifestyles, is also more common in middle-aged and elderly individuals.

Older adults are at a higher risk for respiratory infections (such as pneumonia) and heart disease. Additionally, cancer incidence increases with age, as does the risk of Alzheimer's disease and other age-related cognitive decline.

In summary, age is a crucial factor for understanding the epidemiology of various diseases. Different age groups are more susceptible to different diseases and have different risk factors. Understanding these differences can help public health officials develop more targeted interventions to prevent and treat diseases.

### 2.1.2 Sex/Gender

Disease distribution varies by another crucial epidemiological factor, which is sex or gender. Understanding this attribute can provide valuable insights into the disease status of different populations. Therefore, sex or gender is an essential variable that should be considered when conducting epidemiological surveys and analyzing health data. It is important to note that sex and gender are not interchangeable terms. Sex refers to the biological and physiological characteristics that distinguish males from females, whereas gender is a socially constructed concept that encompasses roles, behaviors, and expectations associated with being male or female.

The prevalence of certain diseases or conditions can be influenced by sex or gender. For instance, breast cancer is more prevalent in women, possibly due to biological factors such as hormones and environmental factors such as exposure to harmful chemicals. On the other hand, some diseases are specific to respective genders because of sex-defining tissues or organs, such as cervical cancer in women and prostate cancer in men.

Mortality rates also differ by gender, with males generally having higher all-cause mortality rates than females. Additionally, certain diseases may affect women

more severely than men. The differences in disease occurrence between sexes can be attributed to several factors, including biological factors, working conditions, and lifestyle factors.

It is important to note that the health conditions of transgender individuals may differ based on their surgery and treatment status. For example, a transgender woman who undergoes hormone therapy may have an increased risk of breast cancer due to hormonal changes. Therefore, it is essential to consider the individual's gender identity and medical history when evaluating their health status.

Understanding the relationship between sex/gender and disease distribution is a critical aspect of descriptive epidemiology. By analyzing the patterns of disease occurrence among different sex or gender groups, researchers and health-care professionals can identify risk factors, develop prevention strategies, and improve treatment outcomes.

### 2.1.3 Race/Ethnicity Differences

The US Census Bureau follows standards set by the US Office of Management and Budget (OMB) to collect and present data on race and ethnicity. The OMB standards identify five minimum race categories: White, Black or African American, American Indian or Alaska Native, Asian, and Native Hawaiian or Other Pacific Islander (Table 1.1). A sixth category, “some Other Race,” is used for people who do not identify with any of the OMB race categories. For ethnicity, the OMB classifies individuals as either “Hispanic or Latino” or “Not Hispanic or Latino.” The Census tabulates statistics on people who report only one race in one of the six “Race Alone” categories and people who report multiple races in the “Multiracial” population [1].

The complex interplay between race and ethnicity as determinants of a disease is an important aspect of descriptive epidemiology. Race is a social construct that categorizes individuals based on their physical characteristics such as skin color,

**Table 1.1** Racial and ethnic categories in the United States

Race	Definition
White	A person of European, Middle Eastern, or North African origin or anyone who identifies as “White” or with specific European ancestry
Black or African American	A person of Black African descent or anyone who identifies as “Black or African American”
American Indian and Alaska Native	A person of Indigenous North or South American origin or anyone who identifies as “American Indian or Alaska Native”
Asian	A person of East, Southeast, or South Asian origin or anyone who identifies as “Asian”
Native Hawaiian and Other Pacific Islander	A person of Indigenous Hawaiian, Samoan, or Pacific Islander origin or anyone who identifies as “Native Hawaiian or Other Pacific Islander”
Two or more races	A person who identifies with more than one of the above races or with a combination of them

whereas ethnicity refers to shared cultural characteristics such as language, religion, and traditions. In the United States, there are significant differences in the prevalence of certain diseases among different racial and ethnic groups. The following are a few examples:

**Sickle Cell Anemia** This is a genetic disease that primarily affects people of African descent as well as those of Hispanic, Middle Eastern, and Mediterranean descent. According to the Centers for Disease Control and Prevention (CDC), sickle cell anemia affects approximately 100,000 Americans, with African Americans being the group most commonly affected [2].

**Skin Cancer** Although skin cancer can affect people of all races and ethnicities, studies have shown that it is more prevalent in Whites compared to Blacks. According to a study published in *Cancer Causes & Control*, non-Hispanic Whites have a higher incidence of melanoma, the deadliest form of skin cancer, than do any other racial or ethnic group in the United States [3].

**Asthma** Asthma is a chronic respiratory condition that can be triggered by environmental factors such as air pollution and allergens. According to a study published in the *Journal of Asthma*, non-Hispanic Black and Puerto Rican children are more likely to have asthma and be hospitalized for asthma than are White children, due in part to a greater exposure to environmental triggers and a lack of access to appropriate medical care [4].

Health disparities are particularly pronounced in racial and ethnic minority populations due to several factors, including socioeconomic status, employment, access to healthcare, and language barriers. For instance, studies have shown that children of certain ethnic groups may have higher rates of asthma due to poor environmental conditions. African Americans and Hispanics are also more likely to experience barriers to health-care access, leading to higher rates of preventable diseases such as hypertension and diabetes.

The complex interplay between race and ethnicity as determinants of a disease underscores the importance of descriptive epidemiology in identifying health disparities and designing targeted interventions to address them. Health policies and interventions must be designed to account for the unique needs of different racial and ethnic groups to ensure that all individuals have equal access to health-care services and resources. This requires a comprehensive approach that involves collaboration between health-care providers, public health officials, and community organizations to promote health equity and eliminate health disparities.

#### 2.1.4 Place-Related Factors

The “where” aspect of disease occurrence is a crucial element in epidemiology, as it helps identify the patterns and potential causes of a disease related to a place. Place can be classified into various discrete units, including, but not limited to, counties, states, regions, and countries. Place of residence, such as urban and rural,

may also influence disease patterns. Understanding the differences in disease occurrence among different places and populations can provide valuable insights into disease risks and its determinants.

In addition to its physical environment, a place can also be described by its biological and social environments, which may play a role in determining the determinants of a disease. For example, the prevalence of certain diseases may be influenced by environmental factors such as temperature, humidity, precipitation, and air pollution. Additionally, social factors such as income, education, and access to healthcare can also impact disease occurrence and spread. For instance, the incidence of dengue fever is often influenced by physical and social factors that promote the breeding of mosquitoes, which carry and transmit the virus. These factors include stagnant water, high humidity, and the low socioeconomic status of a community, which may lack effective vector control mechanisms. Certain tropical regions, such as parts of Asia, Africa, and South America, are particularly vulnerable to dengue outbreaks due to their geography and climate.

The distribution of diseases can differ based on the specific locations of contaminants. The distribution of waterborne diseases can vary depending on the location and source of the water. For example, a study reported from India found that people living near the Ganga River were more likely to contract waterborne diseases such as cholera and typhoid fever due to the high levels of fecal contamination in the river [5]. Similarly, another study conducted in Ghana found that people who lived near water bodies had a higher prevalence of schistosomiasis, a waterborne parasitic infection [6]. Location of industrial areas can also pose a health risk to nearby residents. A study conducted in the Netherlands found that people living near livestock farms had different rates of respiratory morbidity [7]. Similarly, a study from India found that people living near industrial areas had a higher risk of developing respiratory diseases [8].

Skin cancer is a disease that can vary depending on geographic location due to differences in ultraviolet (UV) radiation exposure. The incidence of skin cancer is higher in equatorial regions where the UV index is high compared to the polar regions, where the UV index is usually low. This relationship between skin cancer and latitude can be explained by the amount of UV radiation exposure that individuals receive in different locations. At the equator, the angle of the Sun is more direct, leading to higher levels of UV radiation exposure, which can damage the skin and increase the risk of skin cancer. In contrast, at the poles, the angle of the Sun is less direct, leading to lower levels of UV radiation exposure and a lower incidence of skin cancer.

Urban and rural areas can also differ in terms of disease distribution. Farmers may be more susceptible to cancer due to exposure to agricultural chemicals in rural areas with more agricultural activity, such as farming. A recent study has shown that exposure to certain chemicals such as pesticides is associated with the incidence of cancer in specific locations in the United States [9]. In urban areas, the built environment and social determinants, such as poverty and access to healthcare, can impact disease distribution. For example, research has shown that Black adults in urban areas experience higher age-adjusted mortality rates than do their White counterparts [10].

Place factors are an important aspect of descriptive epidemiology, as they can provide insights into disease occurrence, spread, and risk factors. Understanding the various environmental and social factors that contribute to disease risk in different places can help public health officials develop targeted prevention and control strategies.

### 2.1.5 Time-Related Factors

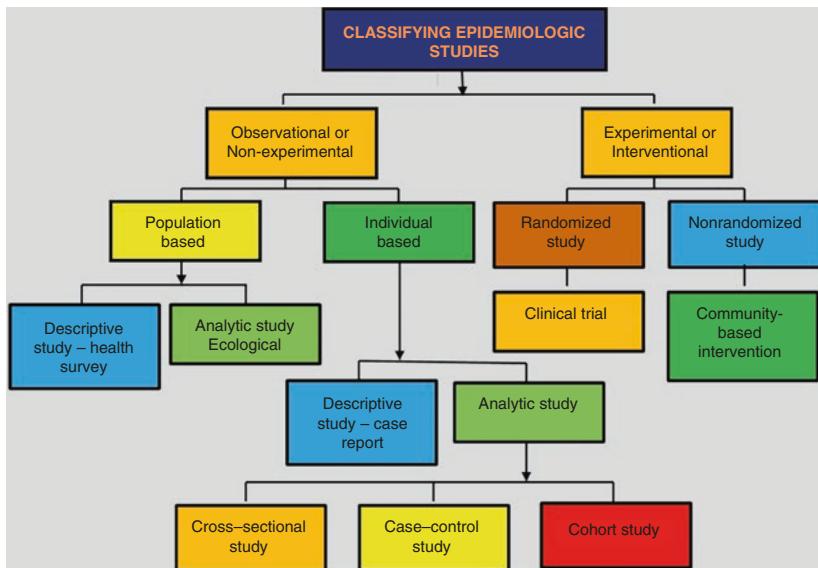
Descriptive epidemiology employs time-related variables to provide a quantitative assessment of the occurrence of health conditions or diseases. Various time intervals such as hours, days, months, and years are used to measure and calculate various health-related events. One crucial concept in this field is temporal trends. These trends focus on changes in the incidence and distribution of diseases over time. Researchers examine temporal trends by analyzing incidence, prevalence, and mortality rates to observe disease patterns over time. For instance, the incidence rate measures new cases of a disease that occur in a population over a specified period. Similarly, the prevalence rate measures the number of individuals who have a particular health condition at a given moment. Prevalence includes both old (existing) and new cases. Prevalence at a given point of time is called point prevalence. Prevalence over a period of time is called period prevalence. The mortality rate determines the number of deaths caused by a particular disease in a given population over a specific time. These measures help epidemiologists assess the temporal trends of various diseases and identify any changes in their patterns over time.

## 3 Descriptive Epidemiological Studies (Box 1.1)

### Box 1.1

Figure 1.2 shows a range of descriptive and analytical studies in epidemiology. Descriptive study designs are well-suited to capturing a snapshot of a health phenomenon and generating hypotheses for future research. These designs offer a valuable perspective on specific health conditions, events, or trends and contribute to the advancement of descriptive epidemiology, which aims to uncover the patterns and distribution of health conditions in populations. Among these study designs, case reports, case series, and cross-sectional studies all play crucial roles in shedding light on the complexities of the health phenomena. By providing detailed descriptions and summaries of health-related issues, these approaches inform public health interventions and policies, ultimately improving the health and well-being of communities

(continued)

**Box 1.1** (continued)


**Fig. 1.2** Types of epidemiological studies. Used with permission from John Wiley & Sons, Inc. from Chapter 17: Investigating the Types of Epidemiologic Studies, Mitra AK, first edition, 2023; permission conveyed through Copyright Clearance Center, Inc. [11]

### 3.1 Case Reports

Case reports are valuable tools in public health research, which provide detailed descriptions of individual health events or conditions. They are commonly used in clinical medicine to capture clinical signs, symptoms, and manifestations of a patient's health condition. These reports often lead to further investigations and discussions of the underlying phenomenon, providing new insights into disease conditions. In fact, some research articles or journals are solely dedicated to summarizing case reports, as they often stimulate further discussion or investigation to understand the underlying phenomenon.

One case report, published in the *American Journal of Tropical Medicine* in 2014, describes the first case of Zika virus infection in a Canadian traveler who had recently visited Thailand [12]. This report emphasizes the importance of considering alternate flavivirus etiologies when dengue serological results are atypical and highlights the need for accurate laboratory testing to differentiate between mosquito-borne infections with overlapping symptoms.

This case report [12] serves as a valuable resource for future investigations into the epidemiology, clinical features, and pathogenesis of Zika virus infection. The report provides a foundation for further research into this emerging infectious

disease by describing one case in detail. Additionally, it underscores the need for improved surveillance and diagnostic tools for mosquito-borne infections in regions where multiple infectious agents co-circulate.

### 3.2 Case Series

A case series is a study design that involves collecting and analyzing a group or series of case reports over time. It can provide important insights into the health phenomena by identifying patterns or trends in a particular disease or condition's occurrence, distribution, and outcomes.

Case series are helpful in situations where a disease's incidence is low or the condition is rare. By aggregating clinical information from multiple cases, epidemiologists can identify clinical course, pathogenesis, and outcomes of a disease or health condition. By aggregating data from multiple cases, researchers can examine similarities in the presentation, progression, and outcomes of the condition.

**Example of a Case Series Study** The study by Gutierrez Sanchez et al. (2022) provides an example of a case series [13]. Their case series of children with acute hepatitis and human adenovirus infection identified a potential association between the two conditions, although the causation remains unclear [13]. The researchers identified 15 children with acute hepatitis, of which 9 were of unknown cause. Eight of the nine children with unknown causes tested positive for human adenovirus. Although liver biopsies did not show evidence of human adenovirus, polymerase chain reaction (PCR) testing of the liver tissue was positive in three children and sequencing of specimens from five children showed three distinct human adenovirus variants. This study suggests that human adenovirus viremia was present in the majority of children with acute hepatitis of unknown cause admitted to Children's of Alabama, but whether human adenovirus was causative remains unclear. The advantage of a case series in this situation is that it allowed researchers to identify a commonality among a group of patients with a rare condition, which could inform future research and clinical guidelines.

### 3.3 Descriptive Cohort (Incidence) Studies

This study design involves following a group of individuals over a period of time to observe the occurrence of a specific health outcome or disease. The primary objective of a descriptive cohort study is to provide a description of the incidence and distribution of the disease or outcome within a population. Descriptive cohort studies are commonly used to generate hypotheses, which can be further tested through analytical cohort studies. A simple example of a descriptive cohort study would involve selecting a group of people in a community and tracking them over time to record the occurrence of a disease, such as diabetes or cardiovascular disease.

It is crucial to note that descriptive cohort studies are different from analytical cohort studies, as the latter is designed to evaluate the relationship between an exposure or risk factor and the occurrence of a disease or outcome through statistical tests. While descriptive cohort studies aim to provide a comprehensive description of the incidence and distribution of a disease or outcome within a population, analytical cohort studies seek to identify the causal effect of specific exposures or risk factors on the occurrence of an outcome or disease.

### ***3.4 Cross-Sectional Studies***

Cross-sectional studies, also called prevalence studies, are designed to measure the exposure and outcome in the study participants for a specific disease or health condition within a specific time frame. Unlike longitudinal studies, which follow participants over time, cross-sectional studies collect data from a single point in time (or over a short period of time) to provide a snapshot of the population's exposure and outcome at a specific time frame.

A cross-sectional study design can be compared to capturing a photograph of a community to obtain a comprehensive understanding of the health or disease burden within the population. Analogous to a photograph, a cross-sectional study captures a single moment in time, providing a static view of the health status of the population. This approach provides valuable insights for epidemiologists into the prevalence of a particular health condition or disease within the community without the need for follow-up observations.

To illustrate, consider a researcher interested in examining the prevalence of smoking within a specific community. Conducting a cross-sectional study would involve recruiting a representative sample of the community and administering a survey to collect data on the participants' smoking habits, demographics, and other relevant factors. The data obtained would provide a snapshot of the smoking behavior of the population at a particular point in time. Another example of a cross-sectional study is determining the prevalence of hypertension among smokers. In this case, researchers would inquire about participants' age, gender, occupation, smoking status, and blood pressure levels to obtain a snapshot of the prevalence of hypertension among smokers.

## **4 Analytical Studies**

Analytical epidemiological studies investigate the relationship between an exposure and outcome in a population. Unlike observational studies, analytical studies are generally more effective in establishing a causal relationship between exposure and outcome. The most commonly used types of analytical studies in epidemiology are ecological studies (correlational studies), case-control studies, cohort studies, and experimental studies.

## 4.1 Ecological Studies (*Correlational Studies*)

One type of analytical study that uses population-level data to investigate the association between an exposure and a disease outcome is ecological or correlational study. In this type of study, data are collected at the group or population level instead of the individual level. Ecological studies are useful in situations where individual-level data are difficult to obtain. They are also called correlational studies because the study design examines the correlation between the two variables to determine whether there is a statistical relationship between them. They are often used to generate hypotheses about the relationship between exposure and disease, which can be tested in more rigorous analytical studies. For example, an ecological study might find a correlation between higher levels of sugar intake in a specific place and higher rates of obesity or diabetes. This suggests that high sugar consumption at the population level may be a risk factor for obesity or diabetes in that population. However, it is important to note that further studies are required to confirm this association.

One advantage of ecological studies is that they are often less expensive and less time-consuming compared to other types of analytical studies, making them an attractive option for researchers. Additionally, ecological studies can provide valuable insights into the relationship between exposures and disease outcomes at the population level. However, ecological studies have limitations that can affect the quality of the findings. One significant disadvantage of ecological studies is the use of population-level data, which lack individual-level information. This issue of using population-based data and making an inference about the association between the exposure and disease without examining the individual-level data is referred to as “ecological fallacy.” This can limit the ability to control for potential confounding variables and establish causality (Table 1.2; Box 1.2).

## 4.2 Case–Control Studies

A case–control study is a type of analytical study that can be conducted relatively quickly and with smaller sample sizes. Such studies are particularly well-suited for investigating rare diseases or analyzing disease outbreaks. In these studies, researchers compare individuals with a specific disease or health condition (cases) to a comparable group without the disease (controls). The control group should be carefully selected to ensure that they are free of the disease in question. The primary measure used in case–control studies is the odds ratio (OR), which is calculated by comparing the odds of the disease in the exposed group to that in the unexposed group.

Health-care professionals often use a  $2 \times 2$  table to organize exposure and disease status data. The odds ratio is calculated as follows (Tables 1.3 and 1.4, Box 1.3):

### Box 1.2

**Table 1.2** Comparison of descriptive and analytical epidemiology

Descriptive epidemiology	Analytical epidemiology
The objective is to describe the distribution of a disease and assess the patterns and trends of disease or health condition	The objective is to identify the causes and risk factors of a disease or health condition
This approach aims to answer questions related to the “who, what, and when” of a health condition or disease	These studies seek to understand the “how and why” of a health condition or disease
The study methods (observational) help generate a hypothesis	The study methods (observational or experimental) help test a hypothesis
Usually, the study methods include an individual or groups to generate a hypothesis	The study methods mostly use study groups for comparison and statistical methods to test a hypothesis
The sources of data are secondary data, vital statistics, government data, and surveys	Based on research questions and hypotheses, new data are collected to test the hypotheses
The study design includes case reports, case series, and cross-sectional studies	The study design includes case-control, cohort, and experimental studies

**Table 1.3** A  $2 \times 2$  contingency table for odds ratio

	Cases (disease +)	Controls (disease -)
Exposed	A	B
Unexposed	C	D
Odds	A/C	B/D

$$\text{Odds ratio (OR)} = AD / BC$$

### Box 1.3

**Table 1.4** Interpretation of odds ratio (OR) values in epidemiological studies

Odds ratio (OR)	Association	Interpretation
OR = 1	No association	The odds of exposure are the same in cases and the control group, indicating no relationship between exposure and disease
OR > 1	Positive association	The odds of exposure are higher in cases compared to the control group, suggesting that the exposure might be a risk factor for the disease
OR < 1	Negative association	The odds of exposure are lower in cases compared to the control group, indicating that the exposure may decrease the risk of developing the disease. In other words, the exposure could potentially act as a protective factor against the disease

### Example of a Case–Control Study

Lam et al. (2013) investigated the association between the use of proton pump inhibitors (PPIs) and histamine 2 receptor antagonists (H2RAs) and vitamin B12 deficiency in a case–control design [14]. This study found that both 2 or more years’ supplies of PPIs and H2RAs were associated with an increased risk for vitamin B12 deficiency, with odds ratios (ORs) of 1.65 (95% confidence interval (95% CI), 1.58–1.73) and 1.25 (95% CI, 1.17–1.34), respectively. The odds ratio for doses more than 1.5 PPI pills/day was even higher at 1.95 (95% CI, 1.77–2.15). This study found that taking more than 1.5 pills per day of PPIs was more strongly associated with vitamin B12 deficiency than taking less than 0.75 pills per day. The odds ratio for vitamin B12 deficiency was 1.95 (95% CI, 1.77–2.15) for doses more than 1.5 PPI pills per day and 1.63 (95% CI, 1.48–1.78) for doses less than 0.75 pills per day.

## 4.3 Cohort Studies

In cohort studies, the researcher selects a group, commonly referred to as a cohort, which does not exhibit any prior occurrence of the outcome of interest or health condition, and then observes the incidence of the outcome of interest in both the exposed and unexposed groups. There are mainly two types of cohort studies: prospective and retrospective.

### 4.3.1 Prospective Cohort Studies

Prospective cohort studies, also known as longitudinal or follow-up studies, involve collecting data on the same group of participants over an extended period. This type of study provides a temporal sequence, which allows researchers to establish a cause-and-effect relationship between the exposure and the outcome. Observing the participants over time allows researchers to determine whether the exposure occurred before the outcome, which is necessary to establish causality.

Once data have been collected in a prospective cohort study, the investigator analyzes them using different statistical methods. These methods can help estimate the risk of developing the outcome of interest related to the exposure of interest. Some commonly used statistical methods include survival analysis and Cox proportional hazards regression.

Large prospective cohort studies have significantly contributed to our understanding of the risk factors for chronic diseases and have played a crucial role in establishing disease prevention and treatment guidelines. The Framingham Heart Study, initiated in 1948, followed a group of participants over several decades to investigate the risk factors for cardiovascular disease. This study has significantly contributed to advancements in the prevention and treatment of heart disease and has played a vital role in establishing guidelines for cardiovascular health [15]. The Nurses’ Health Study, started in 1976, followed a group of female nurses to

investigate the risk factors for chronic diseases such as breast cancer and heart disease [16]. This study provided valuable insights into the impact of lifestyle factors such as diet, exercise, and smoking on health outcomes. Another large prospective cohort study, the Agricultural Health Study, began in 1993 and focused on the health of farmers and their families [17]. This study aimed to investigate the potential health effects of exposure to agricultural chemicals such as pesticides and herbicides. The Framingham Heart Study, the Nurses' Health Study, and the Agricultural Health Study are excellent examples of large prospective cohort studies that have provided valuable insights into the role of lifestyle factors and environmental exposures on health outcomes and have played an instrumental role in guiding public health policies and interventions.

#### 4.3.2 Retrospective Cohort Studies

In this type of study, researchers identify a group of individuals who have been exposed to a particular risk factor in the past and then follow them to observe the outcome of interest. Retrospective cohort studies are also known as historical cohort studies.

Retrospective cohort studies are designed to investigate the relationship between exposure to certain risk factors and the development of chronic diseases or adverse health outcomes. One example of a retrospective cohort study is the study published in the *New England Journal of Medicine* in 2002, which evaluated the association between measles, mumps, and rubella (MMR) vaccination and autism [18]. This study concluded that vaccines are safe and do not cause autism. This study has significantly contributed to our understanding of the relationship between vaccination and autism. Overall, retrospective cohort studies have identified risk factors for many diseases and health outcomes.

### 4.4 Experimental Studies

Experimental studies are crucial for investigating the causal relationships between variables and are considered the gold standard for establishing cause-and-effect relationships. In these studies, researchers manipulate an independent variable while controlling for potential confounding variables and observe its effects on a dependent variable. There are mainly two types of experimental studies: randomized controlled trials (RCTs) and community or field trials.

RCTs are the most common type of experimental study and are often used by clinical researchers to evaluate the effectiveness of new treatments or interventions. Participants are randomly assigned to either an intervention or control group, and the effects of the intervention are measured.

On the other hand, field trials are another type of experimental study that are conducted in real-world settings and are crucial for evaluating the feasibility and effectiveness of interventions in large populations.

One recent example of the application of RCTs in vaccine efficacy and recommendation has come from an article published by Baden et al. (2021) [19]. This study evaluated the efficacy and safety of the Moderna mRNA (messenger RNA) vaccine in preventing coronavirus disease 2019 (COVID-19) in adults. The study used an RCT design, randomly assigning participants to receive either the vaccine or a placebo. The results showed that the vaccine had an efficacy rate of 94.1% in preventing COVID-19 and was generally safe and hence was approved for public use.

In summary, analytical epidemiological studies are essential for understanding the causal relationships between exposures and health outcomes. Case-control studies are particularly valuable for investigating the association between exposure and diseases or outbreaks. Meanwhile, cohort studies are ideal for assessing the impact of risk factors over an extended period. Experimental studies are considered the gold standard for establishing causality, as they allow researchers to manipulate exposure and measure the effect on the outcome of interest. Various statistical tests can help determine the association between exposure and disease. Overall, analytical studies provide a critical framework for identifying and understanding the causes of health outcomes, informing prevention efforts and guiding public health interventions.

## 5 Further Practice

1. Which of the following is a disadvantage of cross-sectional studies?
  - (a) They cannot establish causality.
  - (b) They are costly.
  - (c) They require a large sample size.
  - (d) They are not generalizable to the population.
2. Examples of descriptive epidemiological studies do not usually include:
  - (a) Counts.
  - (b) Cohort studies.
  - (c) Case series.
  - (d) Cross-sectional studies.
3. Ecological studies:
  - (a) Are extremely expensive.
  - (b) May be affected by ecological fallacy.
  - (c) Measure exposure accurately.
  - (d) Yield results that can be applied directly to individuals.

4. Descriptive epidemiology enables the researcher to:
  - (a) Make direct tests of etiological hypotheses.
  - (b) Generate testable hypotheses regarding etiology.
  - (c) All of the above.
  - (d) None of the above.
5. The following study approach is better for rare clinical diseases:
  - (a) Case report
  - (b) Cohort study.
  - (c) Experimental study.
  - (d) Prospective epidemiological study.
6. An investigator wants to study the prevalence of diabetes in a community. Which of the following is the best study design?
  - (a) Case report
  - (b) Case-control study.
  - (c) Cross-sectional study.
  - (d) Community trial.
7. Examples of exposure data in ecological studies include:
  - (a) Income in dollars.
  - (b) Tobacco smoking behavior.
  - (c) Per capita calorie use.
  - (d) All of the above.
8. Case-control studies are among the best observational designs to study diseases of:
  - (a) High prevalence.
  - (b) High validity.
  - (c) Low case fatality.
  - (d) Low prevalence.
9. Which of the following are not person-related factors in descriptive epidemiology?
  - (a) Age.
  - (b) Gender.
  - (c) Marital status.
  - (d) Clustering.
10. Which of the following is a correct interpretation of an odds ratio (OR) value?
  - (a) OR > 1 indicates a negative association between exposure and disease.
  - (b) OR < 1 suggests that the exposure might be a risk factor for the disease.
  - (c) OR = 1 shows a positive association between exposure and disease.
  - (d) OR < 1 indicates that the exposure may decrease the risk of developing the disease.

## Answer Keys

1. (a)
2. (b)
3. (b)
4. (b)
5. (a)
6. (c)
7. (d)
8. (d)
9. (d)
10. (d)

## References

1. U.S. Census Bureau. Measuring racial and ethnic diversity in the 2020 census. 2021. Retrieved from <https://www.census.gov/newsroom/blogs/random-samplings/2021/08/measuring-racial-ethnic-diversity-2020-census.html>
2. Centers for Disease Control and Prevention. Sickle cell disease. 2022. Retrieved from <https://www.cdc.gov/hccbdd/sicklecell/data.html>
3. Garnett E, Townsend J, Steele B, Watson M. Characteristics, rates, and trends of melanoma incidence among Hispanics in the USA. *Cancer Causes Control.* 2016;27:647–59.
4. Oraka E, Iqbal S, Flanders WD, Brinker K, Garbe P. Racial and ethnic disparities in current asthma and emergency department visits: Findings from the National Health Interview Survey, 2001–2010. *J Asthma.* 2013;50(5):488–96.
5. Hamner S, Tripathi A, Mishra RK, Bouskill N, Broadaway SC, Pyle BH, Ford TE. The role of water use patterns and sewage pollution in incidence of water-borne/enteric diseases along the Ganges River in Varanasi, India. *Int J Environ Health Res.* 2006;16(2):113–32.
6. Gyasi SF, Boateng AA, Awuah E, Antwi EO. Elucidating the incidence and the prevalence of Schistosomiasis spp infection in riparian communities of the Bui dam. *J Parasit Dis.* 2019;43:276–88.
7. Smit LA, Hooiveld M, van der Sman-de Beer F, Opstal-van Winden AW, Beekhuizen J, Wouters IM, et al. Air pollution from livestock farms, and asthma, allergic rhinitis and COPD among neighbouring residents. *Occup Environ Med.* 2014;71(2):134–40.
8. Gupta BG, Biswas JK, Agrawal KM. Air pollution from bleaching and dyeing industries creating severe health hazards in Maheshtala textile cluster, West Bengal, India. *Air Soil Water Res.* 2017;10 <https://doi.org/10.1177/1178622117720>.
9. Joseph N, Propper CR, Goebel M, Henry S, Roy I, Kolok AS. Investigation of relationships between the geospatial distribution of cancer incidence and estimated pesticide use in the US West. *GeoHealth.* 2022;6(5):e2021GH000544.
10. Aggarwal R, Chiu N, Loccooh EC, Kazi DS, Yeh RW, Wadhera RK. Rural-urban disparities: diabetes, hypertension, heart disease, and stroke mortality among black and white adults, 1999–2018. *J Am Coll Cardiol.* 2021;77(11):1480–1.
11. Mitra AK. Epidemiology for Dummies. Hoboken, New Jersey, United States: John Wiley & Sons, Inc.; 2023.
12. Fonseca K, Meatherall B, Zarra D, Drebot M, MacDonald J, Pabbajaru K, et al. First case of Zika virus infection in a returning Canadian traveler. *Am J Trop Med Hyg.* 2014;91(5):1035. <https://doi.org/10.4269/ajtmh.14-0151>.

13. Gutierrez Sanchez LH, Shiau H, Baker JM, Saaybi S, Buchfellner M, Britt W, et al. A case series of children with acute hepatitis and human adenovirus infection. *N Engl J Med.* 2022;387(7):620–30.
14. Lam JR, Schneider JL, Zhao W, Corley DA. Proton pump inhibitor and histamine 2 receptor antagonist use and vitamin B12 deficiency. *JAMA.* 2013;310(22):2435–42. <https://doi.org/10.1001/jama.2013.280490>.
15. National Heart, Lung, and Blood Institute. Framingham Heart Study (FHS). n.d.. <https://www.nhlbi.nih.gov/science/framingham-heart-study-fhs>
16. Nurses' Health Study. About the nurses' health study. n.d.. <https://nurseshealthstudy.org/about-nhs>
17. National Institutes of Health. Agricultural Health Study: About. n.d.. <https://aghealth.nih.gov/about/>
18. Madsen KM, Hviid A, Vestergaard M, Schendel D, Wohlfahrt J, Thorsen P, et al. A population-based study of measles, mumps, and rubella vaccination and autism. *N Engl J Med.* 2002;347(19):1477–82.
19. Baden LR, El Sahly HM, Essink B, Kotloff K, Frey S, Novak R, et al, for the COVE Study Group. Efficacy and safety of the mRNA-1273 SARS-CoV-2 vaccine. *N Engl J Med* 2021 ; 384(5), 403–416. <https://doi.org/10.1056/NEJMoa2035389>

# Chapter 2

## Cross-Sectional Study: The Role of Observation in Epidemiological Studies



Jean H. Kim

### Learning Objectives

After completing this chapter, you will be able to:

- Describe how cross-sectional studies are conducted.
- Illustrate the common types of research questions that can be answered.
- Describe the common methods of data collection used in cross-sectional studies.
- Calculate the appropriate measures of both disease frequency and disease association in cross-sectional studies and interpret these statistics.
- Illustrate the role that cross-sectional studies play in epidemiology and public health.

## 1 Introduction

Rudimentary epidemiological studies have been conducted since ancient times in which medical doctors have made conclusions based upon empirical observations of disease patterns. Although the ancient Greek doctor, Hippocrates, lacked the scientific rigor of modern quantitative methods, he was among the first to document his observations that diseases were linked to environmental exposures, personal behaviors, and nutrition [1]. Throughout the centuries, there have been medical observations of the cause of a disease. Modern epidemiology is, however, commonly associated with John Snow, the nineteenth century British physician, who observed a clustering of cholera around homes serviced by a certain water pump in an event that led to the discrediting of the miasma theory of origin [2]. Observations about diseases have led to landmark experimental studies such as the famous study conducted by Dr. Lind to treat scurvy in sailors with citrus rations and cowpox inoculations to prevent smallpox by Dr. Jenner [3, 4].

---

J. H. Kim (✉)

Department of Epidemiology, The Jockey Club School of Public Health and Primary Care,  
The Chinese University of Hong Kong, Shatin, New Territories, Hong Kong  
e-mail: [jhkim@cuhk.edu.hk](mailto:jhkim@cuhk.edu.hk)

Even in contemporary times, developing scientific hypotheses for testing often begins with a simple observation. Often, this observation is made by medical staff and sometimes even by the afflicted populace themselves. In the early twentieth century, it was observed that people with teeth with a discoloration called Colorado brown stain were highly resistant to dental decay [5]. Chemical analysis of drinking water led to the discovery of waterborne fluoride as the cause of these discolorations, which eventually led to widespread fluoridation of municipal water supplies in the United States as a public health measure by the 1960s. This practice then spread to other regions of the world [6].

This chapter will discuss the various types of descriptive studies used in epidemiology, with particular attention to cross-sectional studies.

## 2 Epidemiological Studies: Descriptive Versus Analytical

Epidemiological studies can be divided into descriptive studies, which attempt to characterize disease patterns for hypothesis generation, and etiological (analytical) studies, which examine a research hypothesis of possible disease causation or attempt to quantify the effects [7]. All descriptive studies are observational rather than experimental. Unlike clinical trials and other experimental studies, participants of descriptive studies are not given treatments or other interventions such as health information talks by the researchers or asked to change their behaviors for the study. The individuals included in descriptive studies merely provide information for investigation and are not experimented on in any manner. Table 2.1 illustrates the classification of common study designs used in epidemiological research.

### 2.1 Descriptive Studies: Case Reports/Case Series

A case report is the most rudimentary type of epidemiological study that serves as the first interface between clinical medicine and epidemiology. It simply describes an observed patient with highly unusual symptoms or a medical condition that merits documentation and dissemination to the scientific community. When there is more than one patient, this description of the group of individuals is called a case series. The observations documented in case reports and case series are mainly to alert the health community and raise possible hypotheses for later testing. Often, these hypotheses are tested using a cross-sectional study design. The patient who is usually described often presents with unusual symptoms in which case the case report/case series may be heralding a new disease. The discovery of HIV/AIDS in the 1980s through a case series that noted a clustering of pneumonia due to *Pneumocystis carinii* among young homosexuals in Los Angeles is an important

**Table 2.1** Types of study designs and examples of research questions that may be answered

Type of study	Methodology	Study design	Example of research questions that can be answered
Descriptive studies	Observational	Case report/case series	What are the unexpected side effects or therapeutic benefits of medications?
	Observational	Correlational/ecological study	What is the incidence of breast cancer across countries? Is national stomach cancer prevalence related to the average intake of fish?
	Observational	Cross-sectional survey	What are the seasonal or geographic patterns of suicide in Europe? What age group has the highest level of household injuries? Is there an association between depression and drinking levels in a population?
Analytical (etiological) studies	Observational	Cohort (longitudinal) studies	Does childhood ultraviolet (UV) exposure levels predict skin cancer in later life?
	Observational	Case-control studies	Are workers diagnosed with lymphomas more likely to have been exposed to workplace carcinogens?
	Experimental	Randomized clinical trials/community trials	Is acupuncture effective in alleviating carpal tunnel syndrome versus standard treatment with medication?

historical example [8]. Other examples include severe acute respiratory syndrome (SARS) in 2003, and various other emerging infectious diseases such as Middle East respiratory syndrome (MERS) were normally announced to the world through a case report of the index case or a case series on a cluster of patients [9–12]. As such, these types of studies are highly useful for researchers in the field of emerging infectious diseases and as a surveillance tool for novel disease-causing pathogens. Case reports are a common type of study in disciplines such as alternative and complementary medicine. These studies may report patients that manifest a heretofore unknown adverse reaction to an agent or even poisonings from self-medication with unregulated supplements [13–16]. In the early 1990s, rheumatologist case reports documented cases of a rare disease, eosinophilia–myalgia, leading to further studies conducted by the US Centers for Disease Control and Prevention (CDC) to study over-the-counter supplements as a likely causative agent [17, 18]. Later analytical studies that followed these case reports confirmed this link, leading to the recall of L-tryptophan from an overseas manufacturer after which this disease once again became a rare occurrence [19]. Hence, these types of studies, though simple, play a highly important sentinel role in public health. Case reports/case series can, therefore, serve as alerts for dangerous supplements or even help identify rare adverse events from medications [20].

On the other hand, case reports and case series are also used to alert the medical community of potentially beneficial treatments that were opportunistically discovered by medical staff. There are many examples in which existing drugs were found to have therapeutic effects on unrelated health conditions through such studies. The cosmetic benefits of glaucoma eye drops on promoting eyelash growth was noted from case reports [21, 22]. Similarly, in traditional Chinese medicine, case reports may inform practitioners of the unknown therapeutic benefits of herbal decoctions for treating conditions such as infertility [23, 24].

Case series also can be used to identify hazards of not only treatments but also occupational hazards (risk of injury or risk of health effects from exposure to various factors such as chemicals) [25–28]. Although case reports and case series represent an important study design that may be the first notification of potential hazards, therapies, and diseases, they are extremely limited in terms of providing conclusive evidence of a causal relationship. Since case reports document a single individual and case series rarely document more than several patients, meaningful statistical analysis in these contexts is rarely feasible. Additionally, the findings from a single case or an extremely small number of cases may not be generalizable to a population as a whole. The single individual on whom a case report is based may be an extremely unusual patient with a rare genetic disposition, a highly unusual environment, or rare behaviors that may be unlikely to occur in the general population. For more generalizable findings, a population-based study would be required. Due to the limitations mentioned, case reports/case series cannot be used to make conclusive judgments about the etiology of a health condition without confirmation with other study designs.

## ***2.2 Descriptive Studies: Correlational Studies as the First Step Toward Etiological Examination***

In order to examine the possible etiological factors for a disease, it is helpful to examine the levels of the disease and possible exposures in the general population. Often, epidemiologists may use existing data for these purposes before conducting their own data collection to these ends. One method to quickly obtain information on whether an exposure and an outcome may be causally related would be conducting correlational studies, also known as ecological studies, across different populations. A correlational study involves collecting the entire population's disease frequency measures (e.g., prevalence, incidence, mortality rates). The aggregate characteristic of an entire population, such as a country's breast cancer mortality rate, is the unit of analysis for these studies, which can be compared over time or with other populations [29]. In correlational studies, researchers can examine this disease frequency measure in relation to a possible exposure of interest such as per capita cigarette sales, average levels of alcohol consumption in the population, or

mean consumption of a food or nutrient. For example, the observation that Eskimos had extremely low rates of heart diseases led to the conjecture that their unusual diet may be a protective factor [30, 31]. An ecological study was conducted to examine the rate of ischemic heart disease and stroke mortality in various national populations in relation to average fish consumption, which showed a strong negative correlation [32]. The compelling findings of this correlational study, which suggested that a dietary constituent was the protective factor, was later supported by analytical studies that led to dietary recommendations to consume sufficient levels of omega-3 fatty acids [33].

Although correlational studies provide a useful overview of cross-country and cross-decade time trends and can provide strong clues for disease etiologies, they are nonetheless considered descriptive studies that require follow-up with other study designs. Correlational studies are often a first pass glance to glean clues about disease causation. The main limitation of correlational studies is that exposures and disease cannot be linked at the individual level as these measures are aggregate measures of the entire population (e.g., mean age at the time of marriage by country). Using the above example of the Greenland Eskimos, we cannot link mortality from coronary heart disease to lower fish intake levels at the individual level. It is also possible that the Greenland Eskimos who had higher fish intake may have been more likely to die from heart disease. The data cannot discern these two opposing scenarios. Second, since correlational studies use aggregated data from the population of interest, these studies are prone to what is called “ecological fallacy.” For instance, a country’s average intake of alcohol per capita is typically calculated from the total sales of beer, wine, and spirits divided by the adult population of that country in correlational studies. A country with a high average of alcohol intake may not reflect the drinking levels of a typical person in that population. It would be an ecological fallacy to claim that a typical individual possesses aggregate-level characteristics. It is possible that a small subset of the population has extremely high consumption levels, thereby raising the per capita consumption levels. Alternatively, it is possible that most individuals drink small amounts and that there are virtually no alcohol abstainers in the population, which could also result in similar per capita alcohol consumption levels. These alternative scenarios cannot be discerned from the aggregated data alone. Moreover, the inability to control for possible confounding factors (e.g., other health habits such as exercise levels, risk behaviors such as smoking, or socioeconomic variables such as income) is another major limitation of these correlational designs.

The main advantage of correlational studies, which is the reason why they are commonly used as the first step in investigating an exposure-disease hypothesis, is that they can usually be conducted very rapidly using existing government data, sometimes in a matter of weeks, as opposed to studies requiring several years. As such, these studies are cost-effective ways to explore potential hypotheses before embarking in more time-consuming and resource-intensive studies.

## 2.3 Descriptive Studies: Cross-Sectional Studies: What Are They?

A cross-sectional study, also known as a prevalence study, is a type of descriptive study that addresses many of the shortcomings of case reports/case series and correlational studies. Cross-sectional studies also differ from longitudinal studies in which a study sample is followed over time to record the occurrence of disease in groups that differ by exposure levels (e.g., the incidence or new lung cancer cases among smokers versus nonsmokers over a 20-year follow-up). By contrast, in cross-sectional studies, exposure status and outcomes are simultaneously measured at a single time point. Therefore, the study sample captures people with existing disease (prevalent cases) who are not followed over time. Exposure status and disease status are measured at the same time point. This time point can be a time period that is the same calendar period for all sampled individuals recruited into the study (e.g., January 1, 2022) or it can be variable in calendar date (e.g., 1 month post-hospital discharge, at university entrance, or a survey of teenagers on their 18th birthday). A cross-sectional study records the point prevalence of exposure and health outcomes in this “snapshot of time.” Unlike case reports/case series conducted on a small numbers of patients that occur opportunistically, cross-sectional studies will recruit sufficient study participants for multivariable statistical analyses. Unlike correlational studies, exposures and outcomes can be linked at the individual level with control for potential confounding variables.

An extremely simple cross-sectional study may seek to determine a health condition’s prevalence in a population of interest. For instance, mental health researchers may be interested in determining the proportion of university students who have depressive symptoms at the commencement of the academic year by collecting random samples of these students. A cross-sectional study may also want to go further and examine the health outcome in relation to another variable. Mental health researchers may wish to examine, for instance, depression as a possible outcome of Internet addiction in youths, by collecting random samples of university students and having these students complete diagnostic instruments for assessing depression and Internet addiction. We would then compare the prevalence of depression among those with and without Internet addiction. We could then further study this association after adjusting for potential confounding factors such as academic achievement and household income.

Although the lay population will typically understand the term “exposures” in terms of environmental exposures that usually involve substances that an individual ingests, inhales, or touches, this is only a partial definition. Exposures can include a wide range of factors that can be genetic, environmental, or behavioral (see Box 2.1 and Table 2.2).

### **Box 2.1: What Is an Exposure?**

An exposure is any factor that can theoretically increase or decrease one's likelihood of a health outcome. As such, these factors include not only environmental factors, such as air particulates, viruses, or noise pollution, but also characteristics that are intrinsic to the person, such as sex, race, and other immutable characteristics such as blood type or other genetically determined characteristics. Exposures can also include less tangible characteristics such as attitudes to vaccination, religious beliefs, personality traits, and sexual orientation, as these factors can influence the likelihood of having a health condition.

**Table 2.2** Categories of risk factors and protective factors and methods of data collection

Types of exposures	Examples of the exposure category	Typical methods of collecting data
Sociodemographic characteristics	Sex, age, household income, education, race, type of housing, place of birth	Questionnaires
Environmental exposures	Air pollutants, chemicals in water, radiation levels, noise levels, amount of residential green space	Questionnaires, chemical analysis of air/water/food samples, laboratory testing of human tissues, microbial testing Use of special equipment, geographic information systems
Knowledge/attitudes/psychological traits and states	Contraception knowledge, attitudes toward vaccination, level of introversion, anxiety levels	Questionnaires
Behaviors	Smoking levels, sleep habits, dietary habits, use of social media platforms, personal protective equipment use, annual health screenings	Questionnaires, clinical records documenting behaviors, public observation of individuals, data collection by proxy
Physical attributes	Height, weight, blood pressure, body composition	Observation, medical examination, computer-generated information (e.g., CT scans)

The methods of determining exposures in cross-sectional studies are the same as those with other study designs. Exposures can be determined by direct observation of study subjects (e.g., clinical examination of patients, weight and height measurement). They can be assessed through laboratory tests such as biochemical assays of blood, chemical analysis of air or water samples, and microbial testing. Use of instruments such as computed tomography (CT) scanning and computer-generated

satellite data can also be used to quantify exposures such as exposure to green space. For exposures that are not readily observable or are impractical to observe, questionnaire-based data collection is typically used to obtain information such as sexual history, personal attitudes, health-related knowledge, dietary practices, and behaviors such as time spent using social media. Questionnaire-based data collection for all epidemiological studies can be self-administered (using the paper and pencil format or online platforms) or interviewer-administered, depending on the questionnaire's complexity and the topic's sensitivity. For examples of data collection methods commonly used in cross-sectional studies (and other epidemiological studies), please see Table 2.2. For individuals who are incapable of providing information for themselves (infants, elderly with advanced dementia), information can be collected by proxy from family members or designated caregivers.

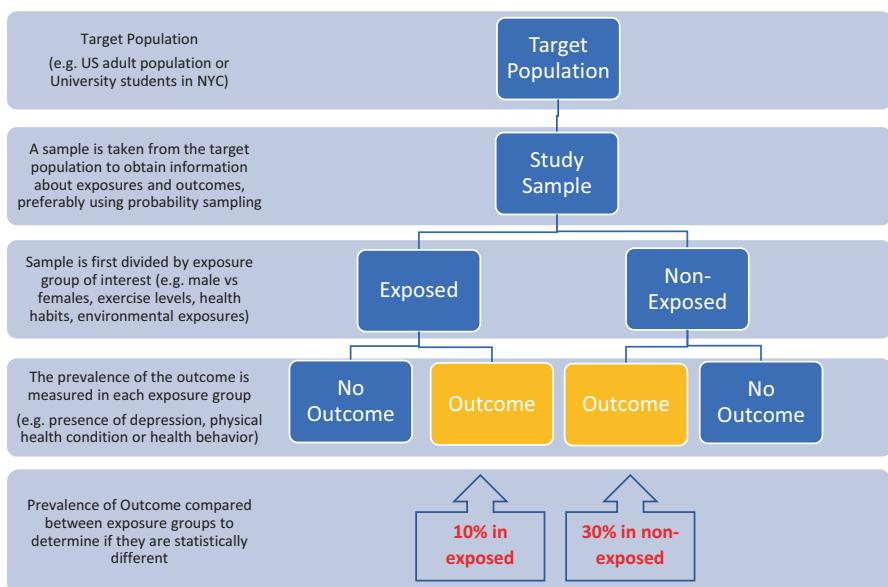
Any method of data collection can be used in epidemiological studies in which the exposures (as well as outcomes) can be enumerated for statistical analysis. However, epidemiological research does not use data collection methods such as focus groups and qualitative interviews. In these qualitative data collection methods, participants are asked to provide subjective interpretations of their life experiences. The data collected using qualitative methods do not use any form of objective measurements or standardized questionnaires with scoring for which statistical methods can be applied. These qualitative data collection methods are sometimes combined with quantitative studies (such as cross-sectional studies), and the combined use of qualitative and quantitative data is called "mixed-method" studies.

In epidemiology, exposures that are hypothesized to cause adverse health outcomes are called risk factors. On the other hand, exposures that are purported to be beneficial for health are called protective factors. For instance, heavy alcohol consumption may be considered a risk for injuries, whereas moderate exercise is a protective factor for many cardiovascular conditions. For the assessment of outcomes of cross-sectional studies, the common methods of assessing health outcomes include data from vital statistics and other government sources of data, hospital records, and occupational and school records. For certain outcomes such as sleep quality, quality of life, and dietary outcomes (e.g., loss of appetite), study participants are often directly asked through self-administered or interviewer-administered questionnaires.

The recruitment of participants into a cross-sectional study requires sampling a group of people from the target population, i.e., the population of which the investigators seek to make inferences about. For instance, investigators may wish to determine the prevalence and risk factors of mental health disorders among elderly in New York City. To obtain a picture of this health condition in a cost-effective manner, studies will typically attempt to assemble a representative sample of elderly rather than attempt to recruit all the elderly residing in New York City. In this sense, cross-sectional studies are like political polls of voter preferences, usually conducted on approximately 1000 respondents before an election. Political polls consist of repeated cross-sectional surveys that track changes in voter preferences over time for the purposes of predicting the likely winner of elections. Unlike cohort studies (longitudinal studies) in which the same group of research subjects are

followed over time, the individuals sampled in these repeated cross-sectional surveys are not the same. Repeated cross-sectional studies are also known as serial cross-sectional studies.

To obtain an accurate picture of the target population, however, the respondents from the target population must be a highly representative sample. A political poll that interviewed only White males or only sampled Native American females, for instance, would be unlikely to have an accurate representation of the voting behaviors of the entire US population. For cross-sectional studies, it is therefore important to use sampling methods that minimize the disparity between the target population of interest and the sampled population. Sampling can be conducted using probability methods (which is scientifically preferred whenever resources and time allow) or non-probability methods such as convenience sampling or snowball sampling. To arrive at correct estimates of the prevalence of exposures or outcomes, it is necessary to have unbiased samples that are as similar as possible to the target population. A noteworthy historical example of biased sampling is exemplified by the 1948 US Presidential election in which political polls based upon telephone surveys erroneously predicted a landslide victory for the Republican candidate, Thomas Dewey [34]. His opponent, Harry Truman, instead won by a landslide vote. Telephones were much more commonly possessed by affluent households who were predominantly Republican, leading to a nonrepresentative sample of American voters.



**Fig. 2.1** An overview of a cross-sectional study design

## 2.4 Cross-Sectional Study Design

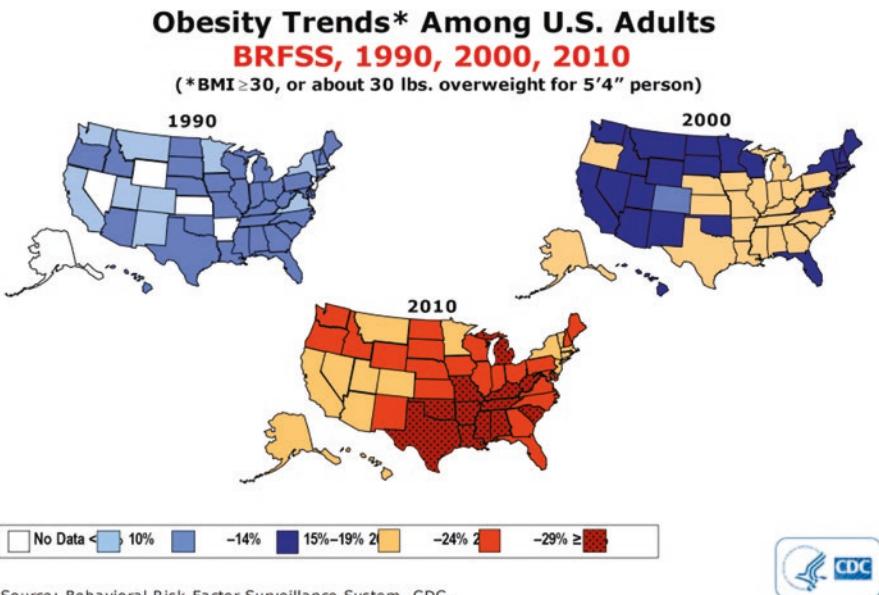
A graphical representation of the conduct of a cross-sectional study is shown in Fig. 2.1 below. As previously described, cross-sectional studies are commonly conducted on a representative sample of individuals from the target population to which we hope to generalize our findings. Once the individuals are recruited and have agreed to participate in the study, we obtain information about exposures (e.g., smoking levels, blood type, dietary patterns, or even information such as political affiliation). We simultaneously obtain or measure the health condition or behavior of interest in each of the participants (e.g., depressive symptoms, cancer diagnosis, and quality of life). It must be noted that, unlike randomized controlled trials, exposures are not assigned by the researcher in a cross-sectional study. Cross-sectional studies are observational, and respondents either have the exposure or do not. The researcher simply collects data on these exposures and outcomes without conducting any experiment or providing interventions/treatments as one of the exposure groups. Once all exposure and outcome information is collected, we can calculate the prevalence of the health condition in each exposure group for comparison. The details of the calculations are shown in a later section of this chapter.

### 2.4.1 What Are Cross-Sectional Studies Used For?

In public health, cross-sectional studies provide insights into the “who, when, where” questions about diseases. For instance, in the example shown below (Fig. 2.2), the prevalence of obesity in US states are shown in 1990, 2000, and 2010 from the prevalence data collected from a series of cross-sectional studies by the US CDC Behavioral Risk Factor Surveillance Survey.

From these data, we can first get an idea of the burden of a disease in the population by simply looking at the proportion of the population with the health condition. Using serial cross-sectionals studies, we can even note population health trends over time and determine whether the burden of disease (e.g., alcoholism or HIV) is increasing or decreasing. It can be easily seen that obesity has been rapidly rising in the United States. The rapid unabated rise of obesity in the United States during this period is useful for health policy agenda setting and informing health promotion policies. In addition to setting priorities for current anti-obesity initiatives such as funding for exercise programs and nutrition education in schools, these data can also be helpful for anticipating and planning future health-care needs. For instance, given the rising trajectory of overweight and obesity in the United States, a higher treatment burden in the future decades (e.g., in 2030) is likely for those diseases for which obesity is a risk factor such as cardiovascular conditions and diabetes. Therefore, cross-sectional studies conducted periodically by the government are an extremely useful data source for health policy planning.

When examining the obesity prevalence in these repeated cross-sectional studies, it is apparent that the patterns differ by geographic region and that there was



**Fig. 2.2** Prevalence of obesity among US adults (in 1990, 2000, and 2010). (Source: <https://www.cdc.gov/obesity/data/prevalence-maps.html#downloads>)

sudden increase in obesity from 1990 onward. These observations from the cross-sectional data may thereby provide clues to the possible reasons for this obesity epidemic. It could be hypothesized, for example, that the timing of the obesity increases corresponded to the rise of the Internet (possibly indicating that a sedentary lifestyle may be partly responsible) as well as the rise of portion sizes in fast-food chains (possibly indicating increased caloric intake as a possible culprit). Geographic variations may provide clues about lifestyle differences, local policy differences (e.g., physical education requirements in schools may differ across different states), or even environmental factors such as climate that may contribute to obesity. For this reason, the United States conducts many periodic cross-sectional studies for assessing the health of its population such as the National Health Interview Survey (NHIS), which has been conducted since 1957 to study topics such as health insurance coverage and immunizations [35].

#### 2.4.2 Examples of Cross-Sectional Studies

The US government also sponsors the National Health and Nutrition Examination Survey (NHANES) that has been conducting periodic surveys on dietary intake and nutrition-related behaviors since 1971 [36]. The US Behavioral Risk Factor Surveillance System (BRFSS) is an example of another important government-sponsored health monitoring system that has been collecting population-level health

statistics. Since 1983, the BRFSS has been conducting annual cross-sectional studies to monitor the prevalence of a wide range of health behaviors such as alcohol consumption, chronic health conditions, and tobacco use [37]. These data allow policymakers to track changes in health behaviors and gauge the effectiveness of policies such as promoting needle exchange and condom use for reducing HIV in the population. These types of studies are useful for informing public health resource allocation.

Another example by which cross-sectional studies can provide useful insights into public health is occupational epidemiology. Cancer prevalence studies by occupation have led to the discovery of occupational risk factors for cancers such as asbestos for mesothelioma, vinyl chlorides for liver cancers, and benzene and its derivatives for lymphatic and hemopoietic cancers [38, 39]. Follow-up studies conducted on these hazards have led to the identification of the causative agents and the development of environmental safety standards. Cross-sectional studies may provide insights into the etiology of health conditions.

Since analytical and experimental studies are typically much more costly and require much more personnel, they are almost always preceded by cross-sectional studies to show that some association exists between the two factors. For instance, when examining the prevalence of mental health problems in relation to marital status, numerous cross-sectional studies have noted a significant association with poor mental health and divorce, and it has been hypothesized that divorce may lead to conditions such as depression and stress. Since the data are obtained on presumed exposures (divorced status) and presumed outcomes (mental health conditions) simultaneously, it cannot be stated unequivocally that divorce caused poor mental health since the temporal sequence of occurrence is unknown. It is often the case in cross-sectional studies that we cannot state unequivocally which of the two factors is the predictor and which is the outcome. Nonetheless, results from cross-sectional studies provide insights by showing that these factors are significantly associated. In order to disentangle the cause-and-effect relationship and show that the presumed exposure indeed preceded the outcome of interest, we would need to conduct longitudinal studies. Cross-sectional studies, nevertheless, are an inexpensive and quick starting point for exploring various hypotheses. Consistent results from cross-sectional studies that show no association need not be followed with longitudinal studies, thereby saving resources.

**Table 2.3** Example of a 2-by-2 contingency table for calculating prevalence by exposure levels

Exposure	Outcomes of interest	
	Binge eater	Non-binge eater
Normal workload ( $\leq 45$ h/week)	223	2490
High workload ( $> 45$ h/week)	201	1098

## 2.5 Analyzing Data from Cross-Sectional Studies

In order to analyze data from cross-sectional studies, the investigators would first examine the difference in prevalence of the health outcome in each exposure category. Let us assume that we researchers are studying binge eating behaviors among American nurses in relation to workload. In order to make inferences about the target population (all practicing nurses in the United States), the researchers would attempt to obtain a random sample of nurses across different regions, ages, and work settings. Once a reasonably representative sample is obtained, the study could ask the nurses to complete a questionnaire about the number of weekly hours they worked in the past month and their eating behaviors. These work hours could be divided into groupings of normal workload ( $\leq 45$  h per week) and high workload ( $> 45$  h per week), and the proportion of nurses in each group that are classified as binge eaters can be ascertained using a contingency table. Below, we show a  $2 \times 2$  contingency table (Table 2.3) representing two levels of exposures and two levels of health outcomes (binge eater or non-binge eater).

We could first compare the prevalence of binge eating in the two groups by calculating the proportion of binge eaters in normal workload nurses versus high workload nurses:

Normal workload nurses:

$$\frac{223}{223 + 2490} = 8.2\% \text{ prevalence}$$

High workload nurses:

$$\frac{201}{201 + 1098} = 15.5\% \text{ prevalence}$$

After calculating the prevalence, we could take the prevalence ratio of the high workload nurses versus the normal workload nurses as the reference group:

$$\frac{15.5\% \text{ prevalence}}{8.2\% \text{ prevalence}} = 1.9 \text{ prevalence ratio}$$

From this, it could be interpreted that those nurses with a high workload have 1.9 times the prevalence of binge eaters. Students new to epidemiology often ask how to decide which group is considered “exposed” and which group is used as the “nonexposed” reference group. In the above example, we have used the normal workload as the reference group. Therefore, the high workload group is considered the exposed group in this analysis. Conversely, if we used the high workload nurses as the reference group, then the ratio would simply be the inverse (with 0.53 as the prevalence ratio). The investigators could also state that nurses with low workload

**Table 2.4** Calculating the odds ratio from contingency tables

Exposure	Outcomes of interest		
	Binge eater	Non-binge eater	Total
High workload nurses (>45 h/week)	a	b	$a + b$
Normal workload nurses ( $\leq 45$ h/week)	c	d	$c + d$

Exposure	Outcomes of interest		
	Binge eater	Non-binge eater	Total
High workload nurses (>45 h/week)	201	1098	1299
Normal workload nurses ( $\leq 45$ h/week)	223	2490	2713

had 0.53 times the prevalence of binge eating compared to those with high working hours.

When comparing these two groups, it is more common to use an odds ratio (OR) for calculating the relative difference between the comparison groups. In tossing a dice with six sides, the probability that the dice lands on a pre-chosen number is 1/6. Unlike prevalence, which is a probability, the term “odds” refers to the ratio of probabilities. The “odds” of the dice landing on a preselected number is the probability of landing on that number (1/6) divided by 1 minus that probability (1-1/6 or 5/6). Odds can therefore be considered as the ratio of the probability of an event over the probability of a nonevent. So, the odds of rolling a dice and having it land on number 2 would be equal to 1/6 divided by 5/6 or 1:5 odds. Using the above  $2 \times 2$  contingency table, we could calculate the odds of binge eating among the two exposure groups (Table 2.4):

Odds of binge eating among high workload nurses:

$$\text{Probability of binge eating status in the high workload group} = \frac{a}{a+b} = \frac{201}{1299}$$

$$\text{Probability of non-binge eating in the high workload group} = \frac{b}{a+b} = \frac{1098}{1299}$$

$$\text{Odds of binge eating in high workload nurses} = \frac{\frac{a}{a+b}}{\frac{b}{a+b}} = \frac{a}{b} = \frac{201}{1098} = 0.183$$

Odds of binge eating among normal workload nurses:

$$\text{Probability of binge eating in the normal workload group} = \frac{c}{c+d} = \frac{223}{2713}$$

$$\text{Probability of non-binge eating in the normal workload group} = \frac{d}{c+d} = \frac{2490}{2713}$$

$$\text{Odds of binge eating in normal workload nurses} = \frac{\frac{c}{d}}{\frac{c+d}{d}} = \frac{c}{c+d} = \frac{0.183}{0.895} = 0.204$$

Taking the ratio of these odds, we would arrive at an odds ratio (OR):

$$\frac{\text{Odds of binge eating in high workload nurses}}{\text{Odds of binge eating in normal workload nurses}} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a}{b} \cdot \frac{d}{c} = \frac{0.183}{0.895} \cdot \frac{0.895}{0.183} = 2.04$$

**Interpretation of Odds Ratio** In comparison with normal workload nurses (the reference group), those in the high workload group had greater than twice the odds of being a binge eater. A closer examination of the above calculations should reveal that the odds ratio can be more simply calculated as the cross product of the cells in a  $2 \times 2$  contingency table as follows:

Calculation of the odds ratio from cross-tabulation of a 2-by-2 contingency table

	Outcome	Non-outcome
Exposed group	A	b
Reference group	C	d

Odds ratio:

$$\frac{a}{c} \div \frac{b}{d} = \frac{ad}{bc}$$

Using the previous example, the odds ratio is simply:

$$\frac{201 * 2490}{223 * 1098} = 2.04$$

An odds ratio of 1.0 indicates that there is no association between the exposure and outcome. Odds ratios can take on any nonnegative numbers from zero to infinity. The further the odds ratio is from 1.0, the stronger the evidence of a causal relationship. However, it is quite rare to encounter odds ratios greater than 10 (indicating an extremely strong risk factor) or less than 0.1 (indicating an extremely strong protective factor).

## 2.6 Advantages and Disadvantages of Cross-Sectional Studies

One major advantage of cross-sectional studies is that these studies can be sometimes conducted with existing data and are thereby much cheaper than longitudinal studies requiring decades of follow-up. An additional advantage of cross-sectional studies is that the investigators can quickly explore multiple possible exposures against multiple possible outcomes inexpensively. Marketing surveys conducted on consumer preferences are examples of commercial applications of cross-sectional surveys that examine various demographic variables (e.g., gender, age, income) with the buying behaviors of many possible products. For public health applications, occupational health researchers, for instance, may be interested in the health effects of various chemicals that workers are exposed to. They may conduct a study on workers asking about exposures to benzene, trichlorides, heavy metals, and whether the worker had experienced health effects such as respiratory problems, dermatitis, nausea, headache, and neurological symptoms. Although we typically cannot conclude definitive causality in finding an association between exposure to one chemical and a reported symptom, these data provide useful information for hypothesis generation that can be confirmed with analytical studies or even laboratory-based research.

Cross-sectional studies are normally considered to be descriptive epidemiological studies rather than etiological/analytical studies that provide evidence of causality. As mentioned earlier in this chapter, the temporal sequence between the presumed exposure and presumed health condition is often not discernible from this “snapshot in time” in which the prevalence of both exposure and health outcomes are captured simultaneously. For instance, a cross-sectional study may note a strong association between unemployment and Internet addiction. It is possible that unemployment led to a much higher likelihood of being an Internet addict (due to greater leisure time). Alternatively, it is also plausible that being an Internet addict (preoccupation with web-based video games, chatting, or web surfing) may have led to unemployment (getting fired from one’s job or possibly never seeking employment due one’s addiction). Due to the single time point of the data collection, it is usually not possible to determine the directionality of even extremely strong significant associations between a presumed exposure and a presumed outcome. Therefore, the causal inferences that can be drawn are limited and must be followed with an analytical study design.

However, when the exposure is fixed in nature (e.g., genetic sex, blood type, race, year, or birth) or is highly unlikely to have resulted from the exposure (place of birth or mother’s age at the time of the respondent’s birth), cross-sectional studies can be considered as analytical studies that provide evidence of a causal

**Table 2.5** An example of misclassification in a cross-sectional study leading to erroneous conclusions

Smoking status	Shortness of breath symptoms	No symptoms	Prevalence of symptoms
Current smoker	21	979	2.1%
Nonsmoker	41	459	8.9%

relationship. For instance, a cross-sectional study of university students shows a statistically significant association between the female sex and anorexia nervosa, and there is no ambiguity about the temporal sequence of these factors. Genetic sex is determined at conception and could not have been caused by the occurrence of an eating disorder in adolescence. In these instances, fixed exposure can be considered as an etiological risk (or protective) factor for the health condition. Suppose that the difference in the prevalence of anorexia nervosa is significantly higher in females than in males. In that case, it can be stated that the female sex is a risk factor for this health condition. Likewise, race is a noted risk factor or protective factor for many health conditions, including major chronic diseases.

As data collection is conducted in a single time point (without follow-up), one additional shortcoming of cross-sectional studies is that misclassification errors can often occur in the classification of exposure or disease, possibly leading to erroneous conclusions. Table 2.5 shows data of a hypothetical cross-sectional study of current smoking status and shortness of breath. Some nonsmokers may be former smokers who have ceased smoking due to respiratory problems. Suppose that the proportion of those who have ceased smoking is high enough among the current nonsmokers. In that case, the prevalence of respiratory symptoms may be higher in nonsmokers than in smokers, leading to the wrong conclusion that smoking is a protective factor for shortness of breath.

If this study were to be conducted using a different way of classifying smoking status (former smoker, current smoker, and never smoker), then we would see the proper relationship between the exposure and outcome.

Smoking status	Shortness of breath symptoms	No symptoms	Prevalence of symptoms
Current smoker	21	979	2.1%
Never smoker	10	340	2.9%
Past smoker	31	119	20.6%

## 2.7 *The Role of Cross-Sectional Studies in Public Health Research and Public Health Policy (Box 2.2)*

### **Box 2.2**

Despite the numerous shortcomings of cross-sectional studies discussed above, these studies are the most common type of epidemiological study due to their cost-effectiveness. Compared with randomized controlled trials, which often cost millions of dollars and require clinically trained staff and onerous documentation of adverse events, cross-sectional studies may be conducted for a fraction of the cost. Since they do not involve following participants over time to observe the incidence of disease or development of symptoms, which may require years or even decades of observation, cross-sectional studies are also highly time-efficient. For small study samples, data collection can be completed in a few days. Hence, during public health emergencies such as the coronavirus disease 2019 (COVID-19) pandemic, infectious disease surveillance researchers will often use cross-sectional studies to quickly determine the prevalence of the disease and possible risk factors for spread. Governments may commission a quick cross-sectional study to determine their populace's current health-care needs to make purchasing decisions and policy decisions based on the most current information. In international health, knowledge, attitudes, and practices (KAP) studies are a type of cross-sectional study commonly employed to quickly characterize the health situation in a new setting.

Due to the speed and low cost of this study design, cross-sectional studies are a popular choice for college and graduate students for their projects. Nonetheless, to establish strong evidence of a causal relationship, cross-sectional studies will often need to be followed with more rigorous study designs, such as prospective cohort studies or case-control studies, which will be covered in the following chapters.

## 2.8 *Problems*

In a cross-sectional study, researchers recruited 1000 ever-married adults from hospital records. It was found that of the adults, 300 were divorced and 700 were currently married. Of the currently divorced respondents, 30 were noted to have clinical depression, whereas 14 of the currently married respondents were noted to be clinically depressed.

Question 1: What was the prevalence of depression in the divorced adults, and what was the prevalence in the currently married respondents? Please calculate the prevalence ratio of depression using married people as the reference group.

Answer:

Prevalence of clinical depression in divorced adults =  $30/300 = 10\%$ .

Prevalence of clinical depression in currently married adults =  $14/700 = 5\%$ .

Prevalence ratio of depression with married adults as the reference group =  $10\%/5\% = 2$ .

Currently, divorced adults had twice the likelihood of clinical depression as compared with currently married adults.

Question 2: By examining the association between divorce and alcohol misuse, explain why a calculated relative risk or odds ratio cannot be taken as evidence of the causal relationship between the two examined factors.

Answer:

In the above scenario, the measure of disease association was not high (only 2.0), and more importantly, the temporal sequence of the relationship between marital status and mental health state is unclear. Marital dissolution may have resulted in poorer mental health. It is also possible that poor mental health contributed to marital dissolution. Additionally, third factors (confounding variables such as income) were not included in the above analysis.

### 3 Further Practice

1. Which of the following is not a major limitation of correlational studies that is overcome by cross-sectional studies?
  - (a) Small datasets.
  - (b) Lack of data on exposures at the individual level.
  - (c) Long period of time needed by investigators to conduct the study.
  - (d) Costliness of correlational studies.
  - (e) None of the above.
2. What is “not” a major advantage of cross-sectional studies as compared with cohort (longitudinal) studies?
  - (a) Cross-sectional studies are generally cheaper.
  - (b) Cross-sectional studies can look at multiple exposures for a health outcome.
  - (c) Cross-sectional studies can be typically conducted in less time than cohort studies.
  - (d) Cross-sectional studies are more representative of the target population.
  - (e) None of the above.

3. Why are cross-sectional studies generally considered descriptive rather than etiological?

- (a) There is no long-term follow-up of respondents.
- (b) We are typically investigating exposures such as biological sex.
- (c) The time sequence of exposure in relation to the outcome is often unclear.
- (d) The study sample is often nonrepresentative of the target population.
- (e) None of the above.

A cross-sectional study recruits 1000 respondents to study the association between daily diet soda consumption and daytime sleepiness. Of these 1000 recruited participants, half consumed diet sodas daily, and 200 of these individuals were observed to suffer from daytime sleepiness. Of those who did not consume diet sodas daily, 100 were classified as having daytime sleepiness.

4. What is the incidence of daytime sleepiness in the diet soda group?

- (a) 50%
- (b) 40%
- (c) 20%
- (d) 10%
- (e) Cannot be calculated.

5. What is the prevalence of daytime sleepiness in the non-diet soda group?

- (a) 10%
- (b) 20%
- (c) 40%
- (d) 50%
- (e) None of the above.

6. What is the odds ratio of the association between daily diet soda consumption and daytime sleepiness?

- (a) 0.17
- (b) 1.0
- (c) 2.0
- (d) 2.67
- (e) None of the above.

7. What is “not” an application of cross-sectional studies?

- (a) Health needs analysis (e.g., determining the number of hospital beds required during a pandemic).
- (b) Projecting health needs in the future (e.g., health-care personnel requirement for chronic diseases).
- (c) Establishing causality between dietary exposures and health outcomes.
- (d) Assessing the likelihood that genetic sex is a risk factor for a disease.
- (e) None of the above.

8. Which is “not” a limitation of case reports/case series that can be overcome by cross-sectional studies?
  - (a) The small sample size of case report/case series.
  - (b) The questionable generalizability of case series findings to the wider population.
  - (c) Lack of detailed information about exposures or outcomes in a case report/case series.
  - (d) Limited statistical analysis that can be conducted on case series data.
  - (e) All of the above are limitations.
9. What is “not true” about serial cross-sectional studies?
  - (a) They can be used for disease surveillance purposes in a population.
  - (b) Serial cross-sectional studies collect data on the same individuals over time.
  - (c) Political polls are examples of serial cross-sectional studies.
  - (d) They can be used to examine changes in disease or exposure prevalence over time.
  - (e) All of the above are false.
10. Which of the following are normative uses of cross-sectional studies?
  - (a) Knowledge–attitude–practices (KAP) surveys.
  - (b) A study to determine the common health problems in a population (e.g., university students).
  - (c) Periodic government-sponsored national health surveys to track risk behaviors.
  - (d) Studies examining demographic factors associated with mental health conditions.
  - (e) All of the above.
11. What are the possible exposures that can be examined in cross-sectional studies?
  - (a) Air pollutants.
  - (b) Genetic mutations.
  - (c) Dietary behaviors.
  - (d) Income.
  - (e) All of the above.
12. Cross-sectional studies can examine the following outcomes except:
  - (a) Chronic diseases such as osteoporosis in the United States.
  - (b) Mental health conditions such as depression among nurses at a hospital.
  - (c) Quality-of-life indicators such as functional status among elderly home residents.
  - (d) Mood states such happiness among infertility patients.
  - (e) Adverse drug events documented in a clinic patient.

## Answer Keys

1. (b)
2. (d)
3. (c)
4. (e)
5. (b)
6. (d)
7. (c)
8. (c)
9. (b)
10. (e)
11. (e)
12. (e)

## References

1. Goldberg H. Hippocrates: father of medicine. 1st ed. Lincoln, NE: iUniverse; 1963.
2. Tulchinsky TH. John snow, cholera, the broad street pump; waterborne diseases then and now. In: Case studies in public health [internet]. Elsevier; 2018. p. 77–99. Available from: <https://linkinghub.elsevier.com/retrieve/pii/B9780128045718000172>.
3. Riedel S. Edward Jenner and the history of smallpox and vaccination. Baylor Univ Med Cent Proc [Internet]. 2005;18(1):21–5. Available from: <https://www.tandfonline.com/doi/full/10.1080/08998280.2005.11928028>
4. Bartholomew M. James Lind's treatise of the scurvy (1753). Postgrad Med J [Internet]. 2002;78(925):695–6. Available from: <https://pmj.bmjjournals.org/lookup/doi/10.1136/pmj.78.925.695>
5. Chandra S. Prevention of oral and dental diseases. Textbook of community dentistry. New Delhi: Jaypee Brothers Medical Publishers Ltd; 2002. p. 107.
6. Howat P, Binns C, Jancey J. New international review supports community water fluoridation as an effective and safe dental health promotion measure. Heal Promot J Aust [Internet]. 2015;26(1):1–3. Available from: [http://doi.wiley.com/10.1071/Hev26n1\\_ED](http://doi.wiley.com/10.1071/Hev26n1_ED)
7. Sackett DL, Rosenberg WM, Gray JA, Haynes RB, Richardson WS. Evidence based medicine: what it is and what it isn't. BMJ [Internet]. 1996;312(7023):71–2. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/8555924>
8. Centers for Disease Control (CDC). Pneumocystis pneumonia—Los Angeles. MMWR Morb Mortal Wkly Rep [Internet]. 1981;30(21):250–2. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/6265753>
9. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus ADME, Fouchier RAM. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med [Internet]. 2012;367(19):1814–20. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23075143>
10. Tsang KW, Ho PL, Ooi GC, Yee WK, Wang T, Chan-Yeung M, et al. A cluster of cases of severe acute respiratory syndrome in Hong Kong. N Engl J Med [Internet]. 2003;348(20):1977–85. Available from: <http://www.nejm.org/doi/10.1056/NEJMoa030666>
11. Chan-Yeung M, Yu WC. Outbreak of severe acute respiratory syndrome in Hong Kong Special Administrative Region: case report. BMJ [Internet]. 2003;326(7394):850–2. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12702616>
12. Centers for Disease Control. Pneumocystis pneumonia—Los Angeles. MMWR. 1981;30(21):1–3.

13. Hsu L-M, Huang Y-S, Tsay S-H, Chang F-Y, Lee S-D. Acute hepatitis induced by Chinese hepatoprotective herb, xiao-chai-hu-tang. *J Chin Med Assoc* [Internet]. 2006;69(2):86–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/16570576>
14. Wasser WG, Feldman NS, D'Agati VD. Chronic renal failure after ingestion of over-the-counter chromium picolinate. *Ann Intern Med* [Internet]. 1997;126(5):410. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/9054292>
15. Huff ML, Fikse D, Surmaitis RM, Greenberg MR. Acute angle closure glaucoma precipitated by homeopathic eyedrops containing Atropa belladonna. *Am J Emerg Med* [Internet]. 2022;54:329.e1–3. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/34776281>
16. Gillett G, Shivakumar N, James A, Salmon J. Acute severe hyponatremia following use of “detox tea”. *Cureus* [Internet]. 2021;13(3):e14184. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/33936895>
17. Hertzman PA, Blevins WL, Mayer J, Greenfield B, Ting M, Gleich GJ. Association of the eosinophilia-myalgia syndrome with the ingestion of tryptophan. *N Engl J Med* [Internet]. 1990;322(13):869–73. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2314421>
18. Martin RW, Duffy J, Engel AG, Lie JT, Bowles CA, Moyer TP, et al. The clinical spectrum of the eosinophilia-myalgia syndrome associated with L-tryptophan ingestion. Clinical features in 20 patients and aspects of pathophysiology. *Ann Intern Med* [Internet]. 1990;113(2):124–34. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2360751>
19. Slutsker L, Hoesly FC, Miller L, Williams LP, Watson JC, Fleming DW. Eosinophilia-myalgia syndrome associated with exposure to tryptophan from a single manufacturer. *JAMA* [Internet]. 1990;264(2):213–7. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/2355442>
20. Vacchiano V, Frattarulo N, Mancinelli L, Foschi M, Carotenuto A, Scandellari C, et al. Teeth loss after teriflunomide treatment: casual or causal? A short case series. *Mult Scler Relat Disord* [Internet]. 2018;24:120–2. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/29982109>
21. Hempstead N, Hempstead RW. Unilateral trichomegaly induced by bimatoprost ophthalmic solution. *J Drugs Dermatol* [Internet], Available from. 3(5):571–2. <http://www.ncbi.nlm.nih.gov/pubmed/15552614>
22. Tosti A, Pazzaglia M, Voudouris S, Tosti G. Hypertrichosis of the eyelashes caused by bimatoprost. *J Am Acad Dermatol* [Internet]. 2004;51(5 Suppl):S149–50. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/15577756>
23. Chao S-L, Huang L-W, Yen H-R. Pregnancy in premature ovarian failure after therapy using Chinese herbal medicine. *Chang Gung Med J* [Internet]. 2003;26(6):449–52. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/12956293>
24. Teng B, Peng J, Ong M, Qu X. Successful pregnancy after treatment with Chinese herbal medicine in a 43-year-old woman with diminished ovarian reserve and multiple uterus fibrosis: a case report. *Med (Basel, Switzerland)* [Internet]. 2017;4(1). Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28930223>
25. Peckham T, Kopstein M, Klein J, Dahlgren J. Benzene-contaminated toluene and acute myeloid leukemia: a case series and review of literature. *Toxicol Ind Health* [Internet]. 2014;30(1):73–81. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22740617>
26. Leikin JB, Carlson A, Rubin R, Secrest CL, Vogel S. Association of Peyronie's disease following petrochemical exposure—a case series. *J Occup Environ Med* [Internet]. 2002;44(2):105–6. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/11851210>
27. Aksoy M. Chronic lymphoid leukaemia and hairy cell leukaemia due to chronic exposure to benzene: report of three cases. *Br J Haematol* [Internet]. 1987;66(2):209–11. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/3606957>
28. Cámará-Lemarroy CR, Gómez-Moreno EI, Rodríguez-Gutiérrez R, González-González JG. Clinical presentation and management in acute toluene intoxication: a case series. *Inhal Toxicol* [Internet]. 2012;24(7):434–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/22642292>
29. Krasovsky K. Alcohol-related mortality in Ukraine. *Drug Alcohol Rev* [Internet]. 2009;28(4):396–405. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/19594794>

30. Kromann N, Green A. Epidemiological studies in the Upernivik District, Greenland. *Acta Med Scand* [Internet]. 2009;208(1–6):401–6. Available from: <https://onlinelibrary.wiley.com/doi/10.1111/j.0954-6820.1980.tb01221.x>
31. Bang HO, Dyerberg J, Hjørne N. The composition of food consumed by Greenland Eskimos. *Acta Med Scand* [Internet]. 2009;200(1–6):69–73. Available from: <https://onlinelibrary.wiley.com/doi/10.1111/j.0954-6820.1976.tb08198.x>
32. Zhang J, Sasaki S, Amano K, Kesteloot H. Fish consumption and mortality from all causes, ischemic heart disease, and stroke: an ecological study. *Prev Med (Baltimore)* [Internet]. 1999;28(5):520–9. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/10329343>
33. National Institutes of Health. Omega-3 Fatty Acids [Internet]. Dietary Supplement Fact Sheets. 2021 [cited 2021 Mar 31]. Available from: <https://ods.od.nih.gov/factsheets/Omega3FattyAcids-HealthProfessional/>
34. Mosteller F. Why did Dewey beat Truman in the pre-election polls of 1948? In: The pleasures of statistics [internet]. New York, NY: Springer New York; 2010. p. 5–17. Available from: [http://link.springer.com/10.1007/978-0-387-77956-0\\_1](http://link.springer.com/10.1007/978-0-387-77956-0_1).
35. Centers for Disease Control (CDC). National Health Interview Survey [Internet]. National Center for Health Statistics. 2022 [cited 2022 Apr 1]. Available from: <https://www.cdc.gov/nchs/nhis/index.htm>
36. CentersforDiseaseControl(CDC).NationalHealthandNutritionExaminationSurvey(NHANES) [Internet]. US Department of Health and Human Services. 2022 [cited 2022 Apr 1]. Available from: <https://health.gov/healthypeople/objectives-and-data/data-sources-and-methods/data-sources/national-health-and-nutrition-examination-survey-nhanes>
37. Centers for Disease Control (CDC). Behavioral risk factor surveillance system. U.S. Department of Health & Human Services; 2022.
38. Giarelli L, Bianchi C, Grandi G. Malignant mesothelioma of the pleura in Trieste, Italy. *Am J Ind Med* [Internet]. 1992;22(4):521–30. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/1442787>
39. Lopez V, Chamoux A, Tempier M, Thiel H, Ughetto S, Trousselard M, et al. The long-term effects of occupational exposure to vinyl chloride monomer on microcirculation: a cross-sectional study 15 years after retirement. *BMJ Open* [Internet]. 2013;3(6) Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23794583>

# Chapter 3

## Case–Control Study



Noraini Abdul Ghafar

### Learning Objectives

After completing this chapter, you will be able to:

- Describe how case–control study is conducted.
- Understand the issues in selecting controls.
- Describe the advantages and disadvantages of case–control study.
- Explain nested case–control study.
- Calculate the odds ratio with 95% confidence intervals.

## 1 Introduction

Case–control study is one of the epidemiological study designs that falls under non-experimental or observational study. This kind of study permits the researcher to determine if an exposure is associated with an outcome. Based on the study design, cases (group known to have the outcome or the disease of interest) and controls (group known to be free from the outcome or the disease of interest) will be determined. Researchers then will look back to find the exposure in each group. Each group’s exposure level is then assessed according to its prevalence. It is reasonable to deduce that exposure may be linked to either an increased or decreased frequency of the outcome of interest if the prevalence of exposure differs between cases and controls [1, 2].

---

N. A. Ghafar (✉)

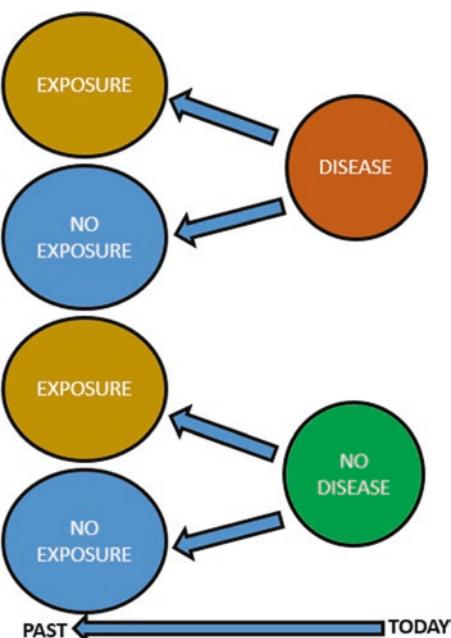
School of Health Sciences, Universiti Sains Malaysia (Health Campus),  
Kubang Kerian, Kelantan, Malaysia  
e-mail: [norainiag@usm.my](mailto:norainiag@usm.my)

## 2 Method of Case–Control Studies

Case–control study is by default a retrospective in nature. However, a special type of case–control study, known as nested case–control study (discussed later in this chapter), is a hybrid design (partly prospective and partly retrospective). In traditional case–control study, cases and controls are selected at the starting point of the study – cases having the disease of interest, and control not having the diseases of interest. Then, researchers go back to collect information about the exposure status for both cases and controls. The association between the exposure and outcome is calculated after getting the exposure data of the participants [3, 4]. The method of case–control study is shown in Fig. 3.1 [5].

Before a case–control study is carefully planned, as with any other study type, the precise hypothesis being investigated must be articulated. Failure to do so may result in poor design and issues with result interpretation. Case–control studies enable the assessment of a variety of exposures that may be connected to a particular disease [3].

**Fig. 3.1** Case–control study design [5]. Used with permission of John Wiley & Sons, Inc. from Chapter 17: Investigating the Types of Epidemiologic Studies, Mitra AK, first edition, 2023; permission conveyed through Copyright Clearance Center, Inc



## 2.1 Selection of Cases

In a case control study, cases must be defined explicitly. Researchers must pay attention to define the cases as precise (and not ambiguous) as possible. For practical purposes, it is frequently useful to divide the diagnostic continuum into “cases” and “non-cases.” Several issues should be considered in determining the cut off for such a division.

For a statistical point of view, the typical practice is to define “normal” as being within two standard deviations of the mean value. This is just an approximate guide to dividing cases from controls.

For the clinical option, certain clinical signs, such as high systolic blood pressure or low glucose tolerance, may be asymptomatic but have a negative prognosis. This may create bias in defining cases and control solely based on clinical parameters. Let us take another example: although normal blood pressure for most adults is defined as a systolic pressure of less than 120 and a diastolic pressure of less than 80, sometimes a male aged 50 or above may have a systolic blood pressure of  $\approx 140$  mm Hg but clinically normal in the absence of symptoms.

Nonetheless, an objective case definition must be used to differentiate cases from control. An investigator should have a distinct goal and purposes for the study before defining cases. Regardless of the method of selection of cases, the case definition should be as explicit and unambiguous as feasible [6].

## 2.2 Selection of Controls

The selection of controls is the next crucial aspect of designing a case–control study. It is important to select controls that are broadly comparable to the cases. The chosen control group needs to be at least similar in the likelihood of developing the result.

It is typical to employ two sources of controls, population control and hospital controls. The advantage of choosing controls from the general population is that their exposures probably represent individuals who are likely to develop cases. Population control is the most desirable method. Non-cases are sampled from the source population giving rise to cases. Another method called neighborhood controls or relative controls is recommended, provided they do not share the exposure of interest, such as smoking in the investigation of cancer.

The use of hospital controls is generally discouraged because of several issues. First, hospital controls may have diseases resulting from the exposure of interest. For example, smoking is an important risk factor for cancer. If we select cases having cancer, the controls may have a disease that is also related to smoking (such as asthma, COPD or heart disease). Secondly, hospital controls may not be representative of the exposure prevalence of the source population of cases. For example, smokers with some illnesses may be more hospitalized than the general population

who are not hospitalized. However, hospital controls are a vital source of controls. These controls are simple to recruit and are more likely to have medical records of comparable quality.

One of the other issues is bias in case-control studies. Cases are more motivated to recall facts of their past events than controls who have no particular interest in the research subject because they are eager to learn what caused their disease [4]. Selecting hospital controls with conditions believed to lead to comparable memory errors may mitigate some of the issues caused by this type of information bias [1]. More information about bias is discussed in Chap. 11.

### **2.3 Issues in Selecting Controls**

Before looking at the issues in control selection, we have to first acknowledge the importance of comparability between cases and control. In case-control study, controls must be comparable with cases. Issues arise when controls were not comparable to cases. Sometimes, selection bias occurs in selecting the controls. The consequence of this is inaccurate results of the analysis.

Four strategies could be used to overcome the problem of selection bias. One of the four strategies may be employed to allow the controls to represent the same population as the cases.

1. A convenience sample – This is one of the most common methods of selecting samples. A convenience sample is drawn like that of the cases, such as by enrolling in the same outpatient department. While undoubtedly convenient, this could weaken the study's external validity.
2. Matching – A matched or unmatched random sample from the unaffected population may serve as the controls. Once more, there are issues with adjusting for unidentified influences, but if the controls are too similar, they might not be representative of the broader population. “Over matching” could lead to an underestimation of the real difference. The benefit of matching is that it enables any given impact to be statistically significant with a smaller sample size.
3. Using a minimum of two control groups – More than one control increases the statistical power of the study. This issue is further discussed later in this chapter. With having more controls, the conclusion is stronger if the study shows a substantial difference between the patients with the desired outcome and those without it, even when the latter group has been sampled in a variety of ways (for example, outpatients, inpatients, and general practice patients).
4. Both patients and controls are drawn from a population-based sample – a random sample of all patients with a particular ailment can be drawn from specified registers. The control group can then be created by choosing randomly selected individuals with similar age and sex distributions from the same population as the area covered by the disease registration.

Meanwhile, researchers can avoid the problem of observation and recall bias by utilizing the blinding technique. A double blinding method is one in which neither the subject nor the observer is aware of their status as a case or control subject. They are also not aware of the study’s main objectives. However, blinding subjects to their case or control status is usually impractical as the subjects already know that they have a disease or illness [3, 7]. Instead, only partial blinding can be done. Asking fictitious questions typically allows one to blind the subjects and observers to the study hypothesis.

## 2.4 *How Many Controls Suitable for Each Case*

Finding a reliable source of cases is typically not too difficult, but choosing controls is more challenging. Controls should ideally meet two criteria. Their exposure to risk factors and confounders should, within the bounds of any matching criteria, be typical of that in the population “at risk” of becoming cases, or those who do not have the disease under research but would be included in the study as cases if they did. The exposures of controls should also be quantifiable with accuracy comparable to that of the cases. It frequently becomes impossible to accomplish both of these goals [4].

When both cases and controls are freely available, it is most efficient to select at least an equal number of each. However, the rarity of the disease being investigated frequently limits the number of instances that may be studied. In this situation, statistical confidence can be strengthened by incorporating multiple controls per cases. There is, however, a law of diminishing returns. Researchers designing case–control studies are typically recommended to include no more than four controls per case because adding additional controls does not add much statistical power by increasing this ratio. Among the factors to be considered when deciding the number of control for a matched case–control are [1] the desired type I error rate, [2] the minimum odds ratio to be detected as statistically significant, [3] the estimated number of cases, [4] the control-to-case ratio in the population, (5) the estimated prevalence of exposure in the control group, and [6] an estimate of the correlation coefficient for exposure between cases and their matched controls [8].

## 2.5 *Advantages and Disadvantages of Case–Control Studies*

Case–control studies are typically rapid, inexpensive, and simple to conduct. Samples of cases and controls are frequently taken from sources like an existing database of patient health records. Additionally, case–control studies are particularly well suited for researching the risk factors linked to uncommon diseases or conditions. In contrast, if the illness or condition is uncommon, an observational design, such as a prospective cohort study, would not be appropriate because it is

unlikely that many participants will experience the illness or condition of interest. Contrary to cohort research, case-control studies are less likely to have loss to follow-up. Before doing more extensive and expensive studies (such as cohort study), case-control studies are sometimes conducted as preliminary research to determine any potential correlations. Case-control studies have the additional drawback of being unsuitable for situations when exposure to any of the risk variables is uncommon because very few, if any, of the cases or controls are likely to have been exposed to them [9].

The biases and interpretation issues that affect all observational epidemiological studies also apply to case-control studies. Confounding, bias in selection or sampling, measurement error, and missing data are a few of these issues. Selection bias is a severe form of bias resulting from missing data in which respondents from the source population who are not included in the study have no observational data at all.

## 2.6 Nested Case-Control Study

The nested case-cohort study is an observational design that incorporates the case-control approach within an established cohort. The design overcomes some of the disadvantages associated with case-control studies while incorporating some of the advantages of a cohort study. For example, in a nested case-control study, the exposure factors such as blood samples for parameters that may determine a disease are already preserved. In this design, researchers start with a suitable cohort that contains a sufficient number of cases (to have sufficient statistical power to address the research topic). The researchers then decide on a random basis a representative sampling of the individuals who have no outcome or the condition being studied (the controls); they pick two or three controls to match with a case. This is done to improve the power of the study [10]. The process of case and control selection is done prospectively in a defined period of time (for example, 5 years). Once cases and controls have been selected, then researchers go back to analyze the already collected samples for laboratory tests.

Using nested case-control studies is a highly effective method for determining the causes of variability in cancer incidence rates within a community. Since a disease-free group of samples within a cohort is selected to begin with, there is less chance of the selection biases that can occur in a conventional case-control study (described earlier), which is solely retrospective. Due to the fact that data gathered as part of a cohort study are collected before the onset of sickness, information bias is less likely to occur. By limiting data extraction and coding to the nested case-control sample, substantial cost savings can frequently be achieved [11].

Other source data may also be collected “retrospectively” on the sampled participants, although the risk of information bias must be taken into account. In comparison to the analysis of the laboratory data for the entire cohort which is typical for a cohort study, a nested case-control sampling does not require such analysis of data of the entire cohort. The increase in efficiency is determined by the number of

controls per case. In many instances, such as examining the association between a disease and a rare exposure, evaluating the impact of confounding, or determining the variance in relative risk with a putative effect modifier, a specified level of efficiency may require a large number of controls [12]. However, a cost-effective analysis is required to select the number of controls per cases [13].

### 3 Example of Case–Control Studies

The following examples of case–control studies have been taken from published sources [14–17]:

**Example 1** Tan et al. (2018) investigated breast cancer risk variables in Malaysian women [14]. Participants in the study are drawn from two hospitals in Selangor, Malaysia: the University Malaya Medical Centre (UMMC), a public hospital, and the Subang Jaya Medical Centre (SJMC), a private hospital. All patients with clinically diagnosed breast cancer were eligible to be included as cases. Cases from UMMC have been recruited since October 2002, while SJMC cases have been recruited from September 2012. Healthy women aged 40 to 74 with no history of breast cancer were recruited for the Malaysian Mammography Study (MyMammo) at UMMC and SJMC. MyMammo at SJMC is a subsidized opportunistic mammography screening initiative that began in 2011. At UMMC, MyMammo began recruitment in 2014 from patients attending normal opportunistic screening. All participants in the study were interviewed by trained interviewers at the hospitals. The participants filled out a questionnaire that asked about their demographics, personal and family history of cancer, history of breast surgery, menstrual and reproductive history, use of oral contraceptives and hormone replacement therapy (HRT), breast cancer diagnosis (cases only), and history of and motivation to attend mammography screening (controls) only. Participants supplied a blood sample, which was processed and stored. After controlling for demographics and other risk variables, participants who had breast surgery to remove cysts and lumps were 2.3 times (95% CI, 1.82 to 2.83) more likely to get breast cancer than those who had never had breast surgery. After controlling for demographics and other risk factors, a first-degree family history of breast cancer was related with a 19% increased risk of breast cancer. After controlling for demographic and other risk variables, “post-menopausal women had a 52% increased risk of breast cancer” [14]. Furthermore, the researchers determined that “breastfeeding, soy consumption, and physical exercise are modifiable risk factors for breast cancer” [14].

**Example 2** This example of case–control study was conducted by Ganesh and colleagues in 2011 [15]. Only male patients were included in the study. Patients were interviewed at Tata Memorial Hospital’s (TMH) outpatient department in Mumbai, India. The data was collected using a predesigned questionnaire that was pre-tested at the hospital. The questionnaire included demographic variables (age, gender, reli-

gion, etc.), lifestyle (habits such as smoking, chewing, drinking alcohol, etc.), dietary habits, and occupational exposure. Patients from all over India come to the hospital because it is a comprehensive cancer clinic for diagnosis and treatment. In general, 30–40% of overall registrations are diagnosed as cancer-free each year. These cancer-free patients were used as “controls” after their medical history and diagnoses were examined. The patients were lung cancer cases that had been microscopically proven. Controls were defined as those that were diagnosed by microscope as “free of cancer” and not having any respiratory tract diseases and therefore diagnosed as “no indication of disease.” Major risk factors that showed a dose-response correlation with lung cancer included: cigarette smoking ( $OR = 5.2$ ), bidi smoking ( $OR = 8.3$ ), and alcohol drinking ( $OR = 1.8$ ). Only red meat consumption indicated a 2.2-fold increased risk among the dietary categories studied. Milk consumption was associated with a 60% reduction in risk, whereas coffee was associated with a twofold increase in risk of lung cancer. Furthermore, pesticide use was linked to a 2.5-fold increased risk of lung cancer [15].

**Example 3** Xi and colleagues (2020) investigated the relationship between maternal lifestyle and the risk of low birth weight in both preterm and term babies [16]. This case-control study was carried out in 14 Chinese hospitals in Jiangmen, Guangdong Province. A stratified sampling strategy was used based on geography. The number of deliveries was used to make a purposeful selection. Hospitals were picked to ensure that each region had at least two hospitals. From August 2015 to May 2016, the patients and controls in this study were recruited from the same hospitals. The researchers found that women who delivered preterm and were physically active (1–3 times per week and 4 times per week, respectively) had a lower risk of having low birth weight babies ( $aOR = 0.584$ , 95% CI = 0.394 to 0.867 and  $aOR = 0.516$ , 95% CI = 0.355 to 0.752). Pregnant women who did not gain enough gestational weight had a higher risk of having low birth weight kids ( $aOR = 2.272$ , 95% CI = 1.626 to 3.176). Women who were exposed to passive smoking had a higher risk of having low birth weight babies ( $aOR = 1.404$ , 95% CI = 1.057 to 1.864). For term deliveries, both insufficient gestational weight increase and excessive gestational weight gain were significantly linked with low birth weight ( $aOR = 1.484$ , 95% CI = 1.103 to 1.998 and  $aOR = 0.369$ , 95% CI = 0.236 to 0.577, respectively). Furthermore, “parity, a history of low birth weight, prenatal treatment, and gestational hypertension were all related with a higher risk of low birth weight” [16].

**Example 4** Shimeles and colleagues (2019) conducted a case-control study to ascertain the risk factors of tuberculosis (TB) in Ethiopia [17]. In the study, the cases were newly detected bacteriologically confirmed pulmonary TB patients aged >15 years, enrolled for treatment in the selected health centers in Addis Ababa. Controls were age- and sex-matched attendees who presented in the same health centers for non-TB health problems. The data collection took place by including all newly registered TB patients until the required sample size was reached. In the study, it was revealed that patients who lived in houses with no window or one win-

dow (suggesting poor ventilation) were almost two times more likely to develop tuberculosis compared to people whose houses had multiple windows ( $aOR = 1.81$ ; 95% CI = 1.06 to 3.07). Besides, previous history of hospital admission was found to pose risk more than three times ( $aOR = 3.39$ ; 95% CI = 1.64 to 7.03). Having a household member who had TB was shown to increase risk of developing TB by threefold ( $aOR = 3.00$ ; 95% CI = 1.60 to 5.62). The study also showed that illiterate TB patients were found to be more than twice more likely to develop TB compared to subjects who can at least read and write ( $aOR, 95\% CI = 2.15, 1.05$  to 4.40). Patients with household income of less than 1000 Birrs (1 Birr = 0.018 US dollar) per month were more than two times more likely to develop TB compared to those who had higher income ( $aOR = 2.2$ ; 95% CI = 1.28 to 3.78). Tobacco use was found as a fourfold risk factor for getting TB ( $aOR = 4.43$ ; 95% CI = 2.10 to 9.3). BCG vaccination, on the other hand, was found to be protective against TB, lowering the risk by one-third ( $aOR = 0.34$ ; 95% CI = 0.22 to 0.54) [17].

## 4 Calculation of Odds Ratio

Case–control studies produce the odds ratio as a measure of the degree of the association between an exposure and the outcome. It is the measure of association that compares the probabilities of disease or an occurrence among those who have been exposed to those who have not (Table 3.1). Its purpose is to establish the association between exposure and outcome [18].

Here,  $A$  = number of exposed subjects and they have the disease;  $B$  = number of exposed subjects but they do not have the disease;  $C$  = number of unexposed subjects and they have the disease;  $D$  = number of unexposed subjects and they do not have the disease.

$$\text{Odds ratio(OR)} = \frac{AD}{BC}$$

OR is a quantitative representation of the strength of association between a cause and an effect when both variables are presented as categorical variables. As a general rule, the greater the OR, the greater the effect on the outcome. OR is interpreted as follows:

**Table 3.1** Calculation of odds ratio for case–control study

	Disease status		
Exposure status	Case	No disease	Total
Yes	$A$	$B$	$A + B$
No	$C$	$D$	$C + D$
Total	$A + C$	$B + D$	$A + B + C + D$

**Table 3.2** Food poisoning and contaminated salad

Category	Cases (food poisoning)	Controls (without food poisoning)	Total
Exposed (ate contaminated salad)	15 (a)	9 (b)	24
Unexposed (did not eat contaminated salad)	8 (c)	32 (d)	40
Total	23	41	64

- OR of 1: There is no difference between the groups; i.e., there would be no association between the exposure (pizza) and the outcome (being ill).
- OR of >1: Suggests that the odds of exposure are positively associated with the adverse outcome compared to the odds of not being exposed.
- OR of <1 Suggests that the odds of exposure are negatively associated with the adverse outcomes compared to the odds of not being exposed [19].

OR is further illustrated by using real-life data in Table 3.2. Based on the table, those who ate the contaminated salad (exposure) were 6.67 times more likely (OR = 6.67) to get food poisoning (outcome), compared to those who did not eat the salad.

$$\text{OR} = \frac{ad}{bc} = \frac{15 \times 32}{9 \times 8} = 6.67$$

#### 4.1 Calculation of 95% Confidence Intervals for OR

Each odds ratio should have a confidence interval (CI) calculated for it. A CI that includes 1.0 indicates that there is no statistically significant correlation between the exposure and the outcome, and that the correlation might have been obtained by chance alone. Without a confidence interval, an odds ratio is not very meaningful [20]. The 95% confidence interval (CI) is used to estimate the precision of the OR [21].

Based on Table 3.2, CI for the OR could be computed by using the following formula:

$$\text{Upper 95\%CI} = e^{\ln(\text{OR}) + 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$$

$$\text{Lower 95\%CI} = e^{\ln(\text{OR}) - 1.96 \sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}}$$

You can also use the following link for a quick calculator: [https://www.medcalc.org/calc/odds\\_ratio.php](https://www.medcalc.org/calc/odds_ratio.php).

Using the formula (or the quick calculator link), the upper limit of 95% CI = 20.70, and the lower limit of 95% CI = 2.15. In other words, the OR in the population from which the sample was drawn varied from 2.15 to 20.70.

## 5 Further Practice

1. Following are the advantages of case control study, except.
  - (a) Multiple exposures or risk factors can be examined.
  - (b) Rates of disease in exposed and unexposed could be determined.
  - (c) Relatively quick to conduct.
  - (d) Can use existing records.
2. Choose the advantage of matching in case–control studies:
  - (a) Decision to match confounding variables is decided at the outset of the study.
  - (b) Requires a matched analysis.
  - (c) Eliminate influence of measurable confounders.
  - (d) Matched variables cannot be examined in the study.
3. Following are techniques that could be used to ensure that the controls to represent the same population as the cases, except.
  - (a) Using a convenience sample.
  - (b) Blinding.
  - (c) Using two or more control groups.
  - (d) Using a population-based sample for both cases and controls.
4. Following is the purpose of matching in case control study, except.
  - (a) To improve study efficiency by improving precision.
  - (b) To enable control in the analysis of unquantifiable factors.
  - (c) To eliminate sampling bias.
  - (d) To make outcome groups comparable on the matching variable.
5. Following are the factors to be considered when deciding the number of control for a matched case control, except.
  - (a) The desired type I error rate.
  - (b) The maximum odds ratio to be detected as statistically significant.
  - (c) The estimated number of cases.
  - (d) The control-to-case ratio in the population.
6. Choose the reason why it is not advisable to use hospital controls.
  - (a) Hospital controls may have diseases resulting from the exposure of interest.

- (b) Hospital controls may be representative of the exposure prevalence of the source population of cases.
  - (c) The exposure of hospital controls may not be comparable with that of cases.
  - (d) Findings from studies using hospital controls tend to overestimate risk because of differential recall.
7. If a control group member had the ailment under investigation, he or she would have been recognized as a prospective case for the study. This is the fundamental notion that must be followed while picking an appropriate control group.

True/False.

8. Adopting hospital-based controls in a case-control study has the advantage of minimizing selection bias.

True/False.

9. A risk ratio or an odds ratio can be calculated in a case-control study.

True/False.

10. If you use more than four controls for each case, the power of the study is much increased.

True/False.

### Answer Keys

1. (d)
2. (c)
3. (b)
4. (c)
5. (b)
6. (a)
7. False
8. True
9. False
10. False.

### References

1. Lewallen S, Courtright P. Epidemiology in practice: case-control studies. *Community Eye Health*. 1998;11(28):57–8. PMID: 17492047; PMCID: PMC1706071
2. dos Santos Silva I. 1999. Chapter 9: case control studies in cancer epidemiology: principles and methods ISBN-13 978-92-832-0405-3.
3. Breslow NE. Statistics in epidemiology: the case-control study. *J Am Stat Assoc*. 1996;91(433):14–28. <https://doi.org/10.1080/01621459.1996.10476660>. PMID: 12155399
4. Setia MS. Methodology series module 2: case control studies. *Indian J Dermatol*. 2016;61:146–51.

5. Mitra AK. Investigating the types of epidemiologic studies. In: Epidemiology for dummies. 1st ed. Hoboken, New Jersey, United States: Wiley; 2023.
6. Critchley J. Epidemiology for the uninitiated, 5th ed. *J Epidemiol Community Health*. 2004;58.
7. Wacholder S, McLaughlin JK, Silverman DT, Mandel JS. Selection of controls in case-control studies. I Principles *Am J Epidemiol*. 1992;135(9):1019–28. <https://doi.org/10.1093/oxfordjournals.aje.a116396>. PMID: 1595688
8. Hennessy S, Bilker WB, Berlin JA, Strom BL. Factors influencing the optimal control-to-case ratio in matched case-control studies. *Am J Epidemiol*. 1999;149(2):195–7. <https://doi.org/10.1093/oxfordjournals.aje.a009786>. Erratum in: *Am J Epidemiol* 1999 Mar 1;149(5):489. PMID: 9921965
9. Sedgwick P. Case-control studies: advantages and disadvantages. *BMJ*. 2014;3(348):f7707. <https://doi.org/10.1136/bmj.f7707>. PMID: 31419845
10. Sedgwick P. Nested case-control studies: advantages and disadvantages. *BMJ* (online). 2014;348:g1532. <https://doi.org/10.1136/bmj.g1532>.
11. Sedgwick P. Nested case-control studies. *BMJ*. 2010;(340):c2582. <https://doi.org/10.1136/bmj.c2582>. PMID: 20484347
12. Ernster. Nested case control studies. *Prev Med*. 1994;23:587–90.
13. Langholz B, Clayton D. Sampling strategies in nested case-control studies. *Environ Health Perspect*. 1994;102 Suppl 8(Suppl 8):47–51. <https://doi.org/10.1289/ehp.94102s847>. PMID: 7851330; PMCID: PMC1566552
14. Tan MM, Ho WK, Yoon SY, Mariapun S, Hasan SN, Lee DS, Hassan T, Lee SY, Phuah SY, Sivanandan K, Ng PP, Rajaram N, Jaganathan M, Jamaris S, Islam T, Rahmat K, Fadzli F, Vijayananthan A, Rajadurai P, See MH, Thong MK, Mohd Taib NA, Yip CH, Teo SH. A case-control study of breast cancer risk factors in 7,663 women in Malaysia. *PLoS One*. 2018;13(9):e0203469. <https://doi.org/10.1371/journal.pone.0203469>. PMID: 30216346; PMCID: PMC6138391
15. Ganesh B, Sushama S, Monika S, Suvarna P. A case-control study of risk factors for lung cancer in Mumbai, India *Asian Pac J Cancer Prev*. 2011;12(2):357–62. PMID: 21545194
16. Xi C, Luo M, Wang T, Wang Y, Wang S, Guo L, Lu C. Association between maternal lifestyle factors and low birth weight in preterm and term births: a case-control study. *Reprod Health*. 2020;17(1):93. <https://doi.org/10.1186/s12978-020-00932-9>.
17. Shimeles E, Enquselassie F, Aseffa A, Tilahun M, Mekonen A, Wondimagegn G, Hailu T. Risk factors for tuberculosis: a case-control study in Addis Ababa, Ethiopia. *PLoS One*. 2019;14(4):e0214235. <https://doi.org/10.1371/journal.pone.0214235>. PMID: 30939169; PMCID: PMC6445425
18. Kalra A. The odds ratio: principles and applications. *J Pract Cardiovasc Sci*. 2016;2:49–51.
19. Szumilas M. Explaining odds ratios. *J Can Acad Child Adolesc Psychiatry*. 2010;19(3):227–9. Erratum in: *J Can Acad Child Adolesc Psychiatry*. 2015 Winter;24(1):58. PMID: 20842279; PMCID: PMC2938757
20. Mann CJ. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emerg Med J*. 2003;20:54–60.
21. Song JW, Chung KC. Observational studies: cohort and case-control studies. *Plast Reconstr Surg.* 2010;126(6):2234–42. <https://doi.org/10.1097/PRS.0b013e3181f44abc>. PMID: 20697313; PMCID: PMC2998589

# Chapter 4

## Cohort Studies



Deepa Valvi and Steven Browning

### Learning Objectives

After completing this chapter, you will be able to:

- Understand the design of a cohort study and differentiate different types of cohort studies.
- Describe key features of conducting cohort studies.
- Review various examples of cohort studies.
- Discuss advantages, disadvantages, and potential biases of cohort studies.

## 1 Introduction

Cohort studies are powerful tools and a suitable choice of study design to conduct research in human populations. Cohort studies are a type of nonexperimental or observational study design. The term *cohort* comes from the Latin word *cohors*, meaning a group of soldiers or a ship's crew [2]. For ancient Roman armies, a cohort was one of ten divisions of a Roman legion, a major military unit. It is a concept of a group of people proceeding together in time [3].

The word “cohort” has been adopted into epidemiology to define a group of people followed up over time. It was first used by W. H. Frost, an epidemiologist in the early 1900s. Frost, in his 1935 publication assessing age-specific mortality rates and tuberculosis, defined the term “cohort” [4]. The modern epidemiological

---

D. Valvi (✉)

Department of Surgery, The University of Kentucky, Lexington, KY, USA  
e-mail: [deepa.valvi@uky.edu](mailto:deepa.valvi@uky.edu)

S. Browning

College of Public Health, The University of Kentucky, Lexington, KY, USA  
e-mail: [srbrown@email.uky.edu](mailto:srbrown@email.uky.edu)

definition of the word now means a “group of people with defined characteristics who are followed up to determine incidence of, or mortality from, some specific disease, all causes of death, or some other outcome” [4]. For example, a “birth cohort” consists of a group of people born during a particular year or the same period that share similar experiences (such as “baby boomers”) (Box 4.1).

### **Box 4.1: Key Features of Cohort Studies**

The key feature of the cohort study design is that participants are followed up over time [1]. Cohort studies are used to study incidence, causes, and prognosis [1]. Cohort studies allow to study the natural history of a disease and to calculate the incidence of disease, absolute and relative risk, risk difference, and attributable risk and hazard ratio.

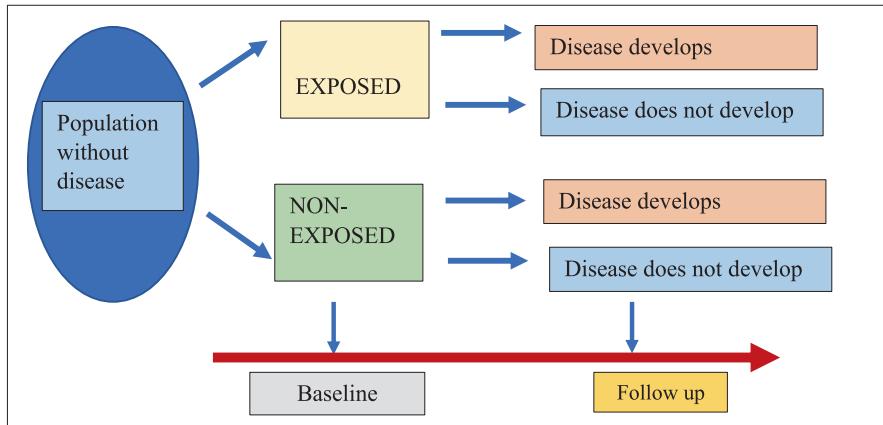
## **2 Design of a Cohort Study**

In the cohort study design, the researcher selects a defined group (the cohort) of healthy participants, classified based on their exposure status, i.e., participants have an exposure (defined as exposed) and others do not have the exposure (defined as nonexposed). At baseline, these participants are disease free, i.e., do not have the outcome of interest to begin with. The researcher follows up both groups (i.e., exposed and nonexposed) over time to determine the occurrence of the outcome of interest or to compare the incidence of disease (or the rate of death from disease) in the two groups (Fig. 4.1). The comparison between exposed and nonexposed groups is the hallmark feature of the cohort study design.

### **2.1 Selection of Study Populations**

As discussed at the beginning of the chapter in the cohort study design, the researcher selects a defined group (the cohort) of healthy participants, classified based on their exposure status, i.e., (A) participants have an exposure (defined as exposed) and (B) others do not have the exposure (defined as nonexposed), i.e., the comparison group. At baseline, these participants are disease-free, i.e., do not have the outcome of interest to begin with.

In an experimental study, the investigators assign participants to the experimental and comparison groups. Cohort studies are similar, but cohort studies are observational, and we do not assign people to smoking (i.e., the exposure) or nonsmoking (nonexposure) since this would be unethical. Instead, cohort studies compare



**Fig. 4.1** Design of a cohort study

people who are smokers in the cohort with those who nonsmokers in the cohort. Cohorts or groups in a cohort study are generated mainly in two ways:

1. The choice of study population is done by selecting groups based on whether they had the exposure or not (e.g., occupational cohort studies), as these occupational workers often have sustained exposures which may not occur in the general population.
2. Second, a cohort can be formed before any of the participants have become exposed or their exposures are unknown. For example, residents from a well-defined geographic area can be taken to study common exposures, such as the Framingham Heart Study.

## 2.2 A Few Points to Consider for Exposure and Comparison Groups

While selecting exposure groups, the investigators must specify the minimum amount of exposure for their particular study. For example, a man may be drinking for a few years, stop for some years, and then again return to drinking after a few years. Thus, investigators can classify the exposure into levels, such as high, medium, or low, to assess the strength of an association (i.e., a dose-response relationship, where the risk of disease increases or decreases as the intensity or duration of the exposure increases or decreases, respectively). Hence, investigators must take into consideration the intensity, duration, regularity, and variability of the exposure [5].

While selecting the nonexposure or the comparison group, it is advisable to choose a comparison group as similar as possible with respect to all other characteristics and factors except the exposure. There are basically two potential sources for

**Table 4.1** Calculation of SMR for uranium workers

Age	National population death rates per 1000	Uranium miner population	Observed death <sup>a</sup>	Expected deaths using the national rates
25–34	4.0	400		1.6
35–44	6.0	500		3.0
45–54	9.0	300		2.7
55–64	25	200		6.0
Total		1400	16	13.3

SMR = Observed deaths × 100 = 16 × 100 = 120

Expected deaths	3.3
-----------------	-----

<sup>a</sup>These values may or may not be known by age group; the total observed death is 16

the comparison group: (1) an internal comparison and (2) the general population. An internal comparison group is choosing individuals from the same cohort who have not been exposed. In an internal comparison group, the relative risk (RR) for exposed and relative risk for the nonexposed are used as the measure of association.

When the general population is the source of the comparison group, one can examine the effects from more than one exposure. Comparison can be done using preexisting data. For example, in the United States, general population data on disease occurrence and death are available from the National Center for Health Statistics, while internationally, the mortality statistics are available from the World Health Organization (WHO). For mortality data, when the general population is the comparison group, the standardized mortality ratio (SMR) is used as a measure of the ratio of the observed deaths with the expected deaths when the rates of the comparison population are used.

**Example** SMR compares the mortality in a study group (e.g., uranium miners) with the mortality that the uranium miners would have had if they had experienced national mortality rates. It gives the likely excess risk of mortality due to uranium mining.

Table 4.1 shows that the observed number of deaths = 16. Expected deaths = 13.3. The calculated SMR = 120. It means that the mortality was 20% higher than that experienced by the national population. Values over 100% represent an increased risk, and those below 100% have a relatively decreased risk.

When population rates are available by age, gender, and race, then SMRs can be adjusted or “standardized” to control for confounding by these factors. Chapter 9 illustrates a detailed method of standardization.

### 3 Types of Cohort Studies

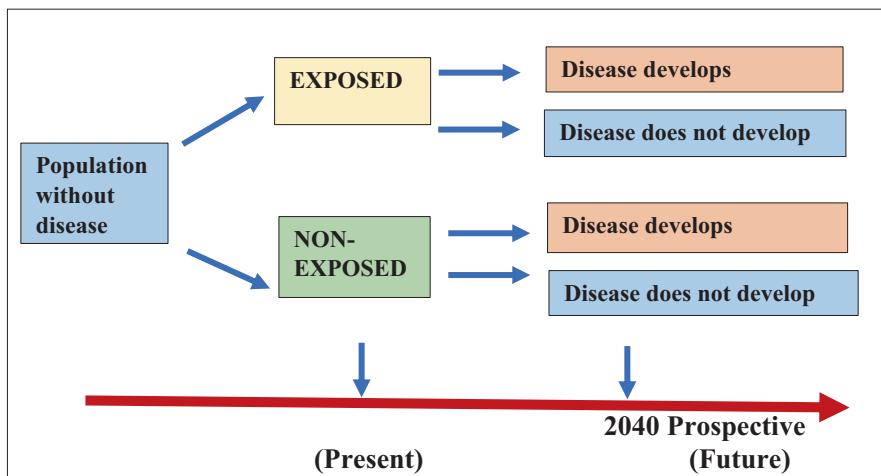
There are basically three different types of cohort studies based on timing of events in a cohort study: (1) prospective cohort study; (2) retrospective cohort study; and (3) ambi-directional cohort study. Let us look at each of these studies in detail:

#### 3.1 Prospective Cohort Studies

In a prospective cohort study, just as the name suggests, the study participants' cohort is formed at the start of the study (i.e., at baseline), based on some past or current exposure (Fig. 4.2). These individuals are followed up over time into the future to ascertain the outcomes of interest. At baseline, or at the start of the study, the outcomes have not yet occurred. The investigator follows up these participants until the point at which the disease develops or disease does not develop to study the outcomes of interest (Fig. 4.2).

The investigator observes these participants, through calendar time, from baseline to looking forward in time, which gives the estimate of the true risk (absolute) for the cohorts being studied. Therefore, it is the gold standard to ascertain absolute risk among observational studies.

**Example** The Nurses' Health Study (NHS) established by Frank Speizer in 1976 is one of the largest prospective cohort studies that investigates the risk factors for major chronic disease in women. The investigators enrolled 121,700 married registered nurses, aged 30–55 in 1976 who lived in the 11 most populous states in the USA. The original focus of the study was on contraceptive methods and its long-



**Fig. 4.2** Prospective cohort study

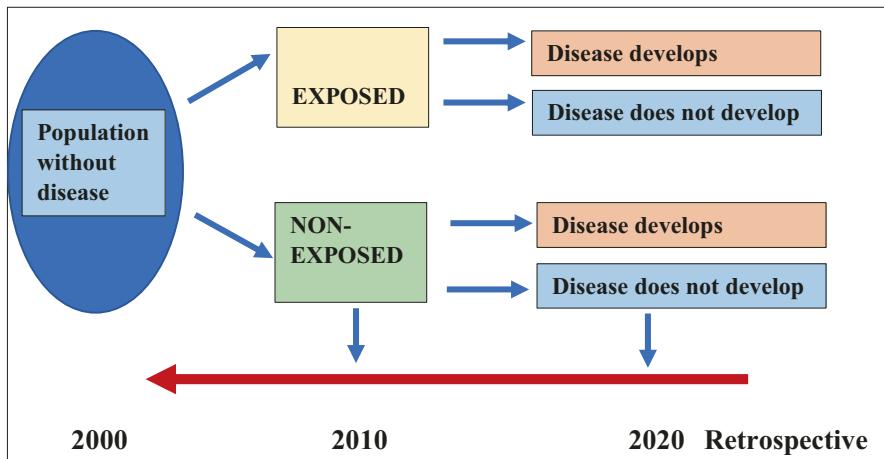
term consequences, smoking, cancer, and heart disease and later studied other life-style factors, behaviors, personal characteristics, and more than 30 diseases. A follow-up questionnaire, every 2 years questioned about diseases and health-related topics, like smoking, hormone use, and menopausal status. A [food-frequency questionnaire](#) was added since 1980, while biological sample collection was added from 1989 to 1990 [6], the NHS II was established in 1989, and NHS III was established in 2010 [6]. Investigators conducted assessments and many significant findings from the NHS that pertain to associations of lifestyles, behaviors, and dietary risk and environment with risk of specific diseases have been reported [7].

The NHS has been going on for more than 45 years. Although prospective studies are expensive, funding has to be continuous, and with such a lengthy duration of a study, the subjects may live longer than the investigators [8], it gives detailed information on exposures and other key variables, as data collection proceeds and follow-up is undertaken [5].

### **3.2 *Retrospective Cohort Studies***

In order to shorten the time taken by a lengthy prospective study, a retrospective cohort study is often considered as an alternative approach. As the name suggests, a retrospective study looks back in time. The data are obtained from previous records; therefore, the exposures and outcomes have already occurred in the past. Even though the outcomes have occurred in the past, the basic study design is the same [9], i.e., comparison of exposed and nonexposed populations (Fig. 4.3). The retrospective cohort design obtains data on outcomes that have occurred and not future outcomes. In other words, a cohort of subjects selected based on exposure status is chosen at the present time, and outcome data (i.e. disease status, event status), which was measured in the past, are reconstructed for analysis [10]. The investigator has limited control of the data quality in this type of study design; however, due to immediate availability of data, this study is cheaper, and it takes a shorter duration compared to prospective studies.

For example, Valvi and colleagues, from 2010 to 2012, examined how fibrinogen levels at baseline affected outcomes of death, development of COPD, lung function decline, and COPD-hospitalizations [11]. In this example, the study population originated from the combined cohorts of the Atherosclerosis Risk in Communities (ARIC), aged 45–64 years from 1987 and Cardiovascular Health Study (CHS), age 65 years, and above from May 1989 to 1990. The exposure here was fibrinogen levels, while the outcomes studied were lung function decline, development of COPD, COPD-hospitalizations, and death. The exposed group had elevated fibrinogen which was defined as  $\geq 393.0$  mg/dL, while the nonexposed group for this analysis was defined as fibrinogen  $<393.0$  mg/dL. The findings revealed that fibrinogen was a significant predictor of all these outcomes with the exception of being in the most rapidly declining FEV1 quartile. Because of available data, another disease,



**Fig. 4.3** Retrospective cohort study

i.e., COPD and its outcomes, could also be studied in a shorter duration and with relatively less expenses.

### 3.3 Ambi-directional Cohort Study

The ambi-directional cohort study as the name suggests combines both the prospective and retrospective aspects. The researcher uses available or existing data to answer a particular research question but, in addition, continues to collect similar data prospectively to address additional research questions. The ambi-directional cohort study design is useful when exposures can cause short-term and long-term outcomes [12].

## 4 Advantages and Disadvantages of Cohort Studies

Population-based cohort studies provide robust results. These studies have made significant contributions to assess risk factor–outcome associations [13]. An outcome or a disease-free study population is first identified by the exposure of interest which is followed over time till the occurrence of the disease or the outcome of interest. They have a temporal framework giving strongest evidence to assess causality and are most useful when compared to RCTs to study certain exposures; when we cannot expose people to asbestos or cigarette smoking as this would be unethical. They are useful to study rare exposures because participation of subjects is

**Table 4.2** Advantages and disadvantages of cohort studies

<b>Advantages</b>
Can assess causality; data is gathered on sequence of events
Examine multiple outcomes for a given exposure
Good to examine rare exposures
Allows calculation of rates of disease in the exposed and unexposed over time (e.g., incidence, relative risk)
<b>Disadvantages</b>
Need to follow large number of subjects to study rare exposures
Vulnerable to selection bias
<b>Prospective cohort studies</b>
May be expensive to conduct and time consuming
May require long durations of follow-up
Not good for diseases with long latency period
Susceptible to loss to follow-up or withdrawals
<b>Retrospective cohort studies</b>
Less control over variables
Susceptible to recall bias or information bias

based on the exposure status. Cohort study also efficient to examine multiple outcomes simultaneously (see Table 4.2).

Disadvantages include the potentially long follow-up times, the need for large sample sizes, and the consequent high costs to conduct the study, therefore diseases such as cancers, cardiovascular diseases, and neurological conditions that have a long latency period are not suited for cohort studies. To conduct a study in young disease-free cigarette smokers and nonsmokers, we would have to wait almost 20 years to monitor some disease occurrence in this population. These studies are also not suitable for rare diseases as a large number of participants would be needed, and it would be difficult to monitor a large number of people for a long time. Due to a long study duration, the exposure to a certain risk factor could change over time, and some participants may choose to withdraw from the study.

## 5 Examples of Cohort Studies

### 5.1 The Framingham Heart Study

The Framingham Heart Study [14] is considered as the most important cohort study on cardiovascular diseases. The town of Framingham is in the state of Massachusetts, USA, about 20 miles from the city of Boston. By 1940s, cardiovascular disease was responsible for almost half of the deaths in America [15]. As legislators drafted the National Heart Act, a young physician and officer Gilcin Meadors, was tasked to compile a proposal for a future epidemiological study [14]. The proposal entitled “to study the expression of coronary artery disease in a ‘normal’ or unselected

**Table 4.3** Population cohorts of the Framingham Heart Study

	First year	Size	Female %	Salient features
Original	1948	5209	55%	
Offspring	1971	5124	52%	Children of the original cohort and their spouses
Third generation	2002	4095	53%	Children of the offspring cohort
New offspring Spouse	2003	103	54%	Spouses of offspring cohort participants who were not initially enrolled in the study, with at least two children in the third-generation cohort; added to improve statistical power
Omni 1	1994	506	58%	To reflect the increasing diversity, participants from African American, Asian, Indian, Pacific islander, and native American ethnic groups
Omni 2	2003	410	57%	Recruited to achieve 10% of third-generation cohort size

*Note.* From “The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective,” by S.S. Mahmood, D. Levy, R.S. Vasan, and T.J. Wang, 2014, *Lancet* 383(9921):999–1008

population and to determine the factors predisposing to the development of the disease through clinical and laboratory exam and long term follow-up” [16] set the tone for the next 70 years.

The original cohort was recruited between 1948 and 1952, and the age range between 28 and 62 years was selected based on the assumption that younger adults would generally be free of the cardiovascular endpoints that were proposed in the 20-year follow-up. The final sample consisted of 5209 participants aged between 30 and 62 years without cardiovascular disease at baseline (Table 4.3). More than half of the participants were women, by contrast with contemporary epidemiological studies, which had very small numbers of women or excluded them altogether [17, 18].

The investigators hypothesized that cardiovascular diseases do not have a single cause but are an accumulation of multiple factors. Some of the hypothesis tested were: does the risk of coronary heart disease increase (CHD) with age? Or does heart disease occur earlier in males? Are tobacco smoking and habitual alcohol consumption associated with CHD?

They identified new coronary events by daily surveillance of hospitalizations in Framingham. A fourfold increase in the incidence of coronary heart disease was found after defining hypertension as blood pressure of 160/95 mm of Hg or higher [19]. Subsequently, high blood pressure was also a notable consequence for stroke [20].

This study selected a defined population based on residence and was not constructed as an investigation in which the cohort was defined in relationship to any single exposure or other risk factor. This design was useful in identifying individuals most likely to have a cardiovascular event based on existing exposure(s) for future outcome(s) of interest. It also proved advantageous to simultaneously study multiple exposures such as smoking, obesity, hypertension, and cholesterol levels.

## 5.2 *Cardiovascular Health Study*

The Cardiovascular Health Study [21] is a population-based study of coronary heart disease (CHD) and stroke in adults 65 years and older. The study was designed to assess the association of risk factors with “subclinical disease” – a change in markers of disease which have not yet manifested as clinical signs of disease. The study recruited a cohort from four US communities: each of the cohort centers recruited 1250 men and women 65 years and older in the first year beginning in June 1989, for a study-wide population of 5000 participants. The study recruited a supplemental cohort of 685 African American men and women in 1992 and 1993 from three of the communities.

The study was designed for the following hypotheses, such as (a) to quantify conventional and novel risk factors of stroke and CHD, (b) to determine the incidence of CHD and stroke with indicators subclinical disease identified by noninvasive procedures like carotid ultrasonography and echocardiography, (c) to assess conventional and novel risk factors are associated with subclinical disease, (d) to understand the natural history of the disease and identify factors associated with clinical course, and (e) to estimate the prevalence of risk factors, subclinical disease, and clinically diagnosed stroke and CHD.

The study conducted extensive examinations on health status, physical activity, physical function, and use of medications. The participants were followed up for 5 years, with 646 reported deaths representing 12% of the population. The study reported that quantitative measures of disease were better predictors of mortality than clinical history of the disease. [22] As with previous studies, the study found male sex, poverty, smoking more than 50 pack-years, lack of physical activity, frailty and disability as predictors of 5-year mortality in older adults. The study helped to contribute to the understanding the influence of the risk factors among previously under-represented groups such as women, older adults, and African Americans, on the risk of cardiovascular diseases.

## 6 Bias in Cohort Studies

Bias is any systematic error that can lead to mistaken results [23]. Biases can occur in any research, and they must be avoided or must be considered while conducting cohort studies. These biases can occur during: (i) the design phase of the study such as selection of sample, or data collection and (ii) analysis phase of the study, i.e., statistical analysis or interpretation of data. Most of the biases have common terminologies, but largely fall into two broad groups: selection bias and information bias. Bias is discussed in detail in another section of the textbook (Chap. 13).

## 6.1 Selection Bias

As the name suggests selection bias occurs when selected study subjects whether exposed or nonexposed give a result among participants that is different from the result that would occur among individuals who were eligible but not included in the study. It is vital to understand that both the exposed group and the nonexposed group are representative of the general population from where they are taken to be able to generalize the results [23]. Appropriate assessment of exposure or an exposed individual is a very important step in the selection of subjects, because inappropriate selection of exposure status will lead to selection bias. Retrospective cohort studies are more predisposed, since exposure and disease have already occurred by the time of subject selection. Health-seeking behaviors for patients differ based on factors such as employment, socioeconomic, and insurance status. Consequently, designing cohort studies which employ methods for sampling that avoid biased selection of participants and high response rates is needed. Selection bias can occur when compared cohorts are part of a population who receive public health intervention, so the exposure can be misled by this influence [23]. Selection bias can also occur due to nonresponders, or no participation bias, but is less frequent in prospective cohort studies as strict follow-up is done [23].

The “healthy-worker effect” is another form of selection bias that occurs in two special types of cohort studies – proportional mortality ratio (PMR) and standardized mortality ratio (SMR) studies. The healthy-worker effect occurs in these studies because the general population, which consists of both healthy and sick people, is selected for comparison to a relatively healthy working population [24].

In cohort studies, loss to follow-up (due to intractability or lack of participation) is a major concern giving rise to selection bias. As cohort studies require long-term follow-up, usually months to years of follow-up, it is expected that due to various life circumstances, participants can get lost during the study, leading to misclassification bias. Suppose, people are lost to follow-up due to a certain disease, then those lost to follow-up differ from those who have not lost to follow-up. The incidence rates of the two will differ, and the results of the analyses will be difficult to interpret.

## 6.2 Information Bias

Information bias can originate in the observed individuals, or in the observers or in the data instruments used to assess the outcomes [23]. Information bias can occur in cohort studies if the information in a study is either measured or recorded inaccurately, i.e., a systematic distortion either in assessing the exposures and diseases. If information collected or measured varies between the exposed and nonexposed persons, this can introduce a significant amount of bias in the study. Therefore, comparable methods for both the exposed and unexposed groups are advisable. Bias can

also occur if the person recording the information is aware of the disease or exposure status of participants under study. Loss to follow-up can lead to information bias due to missing data (i.e., nonresponders or no participation in the study) during the analysis phase of the study. To prevent this bias during the analysis phase of the study, it is possible to exclude individuals who have missing data from the analyses. However, this decision is under the sole judgment of the researchers, if the remaining sample after excluding missing values still gives sufficient statistical power to the study to validate the results [23].

## 7 Measure and Analysis in Cohort Studies

Cohort studies are longitudinal studies where participants belong either to the exposed or the unexposed group and the incidence of the outcome (e.g., disease, death, or change in health status) is observed during the follow-up period.

Incidence risks and rates are the two measures that are calculated from cohort studies. In addition, risks and rates can provide further information on the effects of the exposure of interest, such as risk ratios, rate ratios, attributable risk (risk or rate differences), and attributable risk percent.

### 7.1 Incidence Rate

Risk is defined as the number of new cases in the numerator divided by the total population who were at risk of getting the disease at baseline (start of the follow-up period), i.e., the denominator. Risk is also known as incidence proportion or cumulative incidence. When the denominator is the size of the population at the start of the time period, the measure is called *cumulative incidence*. This measure is a proportion, because all persons in the numerator are also in the denominator and hence also called the incidence proportion. It is a measure of the *probability* or *risk* of disease, i.e., what proportion of the population will develop illness during the specified time period. If all cohort participants were followed up for the same time period, risk can be calculated as the measure of disease occurrence in the exposed groups as well as the nonexposed group.

$$\text{Incidence Proportion} = \frac{\# \text{ of new cases}}{\text{total population at risk}}$$

The other measure for incidence can be the incidence rate, and it is also known as incidence density. Rate is simply as change over time, here, it is a change from

health to disease. The numerator is the number of new cases over a particular time-period divided by the total person-time-at-risk for the population. Sometimes, there can be unequal follow-up periods in a study. For example, if subjects were recruited at different time points into the study, over a long period of time, or if some were lost to follow-up during the study due to various reasons then the follow-up time durations will differ significantly. To handle such situations, one can use person-time at risk (i.e., person-years, person-months, or person-days, etc.) as the denominator (person-time is explained below).

$$\text{Incidence Rate} = \frac{\text{# of new cases}}{\text{total person time at risk}}$$

## 7.2 Calculating Person-Time

As explained above, calculating incidence rate involves determining the amount of person-time accrued by each study subject. Person-time is basically follow-up time of years of exposure.

The denominator, i.e., person-time is calculated by the sum total of time all individuals that remain in the study without developing the outcome of interest. Person-time can be measured in days, months, or years, depending on the unit of time that is relevant to the study.

For example, if a subject enrolled into a study develops the disease under study after 5 years will contribute 5 person-years to the denominator. In case a person is disease free at 1 year and later is lost to follow-up will contribute only one person-year to the denominator. Person-time rates are used in cohort studies (which are follow-up studies) for diseases that have a long incubation or latency periods such as COPD, or occupationally related diseases, AIDS, and chronic diseases.

## 7.3 Risk Ratio (Relative Risk)

	Diseased	Not diseased	Total
Exposed	$a$	$b$	$a + b$
Not exposed	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

Risk ratios can be calculated using two-by-two tables. Risk ratio is the measure that compares the risk in the exposed ( $I_{\text{exposed}}$ ) to the risk in the unexposed ( $I_{\text{unexposed}}$ ). It is defined as the risk in the exposed group divided by the risk in the unexposed group.

Risk ratio is also known as relative risk (RR), computed by:

$$\text{Risk Ratio} = I_{\text{exposed}} / I_{\text{unexposed}} = a / (a + b) \div c / (c + d)$$

For example, in a cohort study, data on exposure to coal mine dust and pneumoconiosis among coal miners was collected. Here, the risk in the exposed is calculated as 65/210 or 0.31 cases per person (31 cases per 100 persons), and the risk in the unexposed is calculated as 25/190 or 0.13 cases per person (13 cases per 100 persons). Risk ratio can be written as 0.31/0.13, or 2.38, which is interpreted as the risk of developing pneumoconiosis is 2.38 times in the exposed compared to the unexposed group.

	Disease	No disease	Total
Exposed	65	145	210
Not exposed	25	165	190
Total	90	310	400

## 7.4 Rate Ratio

Similarly, rate ratio can be computed in a cohort study, as a measure that compares the rate in the exposed ( $IR_{\text{exposed}}$ ) to the rate in the unexposed ( $IR_{\text{unexposed}}$ ). It is defined as the rate in the exposed group divided by the rate in the unexposed group. For the rate ratio, the calculation is based on the person-time contributed to the study.

$$\text{Rate Ratio} = IR_{\text{exposed}} / IR_{\text{unexposed}}$$

From the previous example, if person-time at risk for each coal worker was calculated, then the table would be written as follows:

	Disease	No disease	Person-years at risk
Exposed	65	145	185
Unexposed	25	165	198
Total	90	310	383

In this study, the rate in the exposed will be 65/185 or 0.35 cases per person-years. The rate in the unexposed group is 25/198 or 0.13 cases per person-years. The rate

**Table 4.4** Interpretation of relative risk

Risk ratio (=RR)	Exposure
<1	Exposure is protective
=1	Exposure is neither preventive nor harmful (no association)
>1	Exposure is harmful

ratio can be written as 0.35/0.13, or 2.7, which is higher than the risk ratio calculated above. This can be interpreted as the rate of pneumoconiosis is 2.7 times higher in the exposed as compared to the unexposed group. Table 4.4 can be applied for both risk ratios and rate ratios.

A risk ratio less than 1.0 indicates a decreased risk for the numerator (i.e., the exposed group), than the risk in the nonexposed group. Assuming that there is no confounding or no other factors that can confound the association, we can conclude that there is a protective effect of a factor in the exposed group or the exposure is preventive. A risk ratio of 1.0 indicates identical risk in the two groups, i.e., there is no difference in risk or rates between the exposed and the unexposed groups. Whereas, as risk ratio or greater than 1.0 indicates an increased risk for the exposed group (numerator), therefore the exposure is harmful. The farther the risk ratio is from 1.0, the greater is the effect of the exposure on the study cohort.

## 7.5 Attributable Risk

Measures important from a public health perspective that are helpful in program planning are risk difference and rate difference. The measures of difference are absolute measures, while ratio measures are relative. Risk and rate difference are useful to address questions like “how much of the disease in the exposed group is attributable to the exposure?” Risk difference is also known as attributable risk, as the name suggests it is difference in the risk between the two groups, which indicates the excess risk due to the presence of the exposure. Note that risk difference or attributable risk is a measure of proportion.

$$\begin{aligned}\text{Risk Difference} &= I_{\text{exposed}} - I_{\text{unexposed}} \\ \text{Rate Difference} &= IR_{\text{exposed}} - IR_{\text{unexposed}}\end{aligned}$$

## 8 Further Practice

1. A cohort study is which type of study?
  - (a) Cross-sectional.
  - (b) Experimental.
  - (c) Longitudinal.
  - (d) None of the above.
2. A hallmark feature of the cohort study design is.
  - (a) No comparison between exposed and nonexposed groups.
  - (b) Comparison between exposed and nonexposed groups.
  - (c) Comparison between exposed groups only.
  - (d) Comparison between nonexposed groups only.
3. In selection of study population for cohort studies is which of the following is not true.
  - (a) Study population are assigned to the exposure group or nonexposure group.
  - (b) Selection of study population is not done at baseline.
  - (c) Selection of study population cannot be taken from a well-defined geographic area.
  - (d) Study population are not assigned to the exposure group or nonexposure group.
4. The effects for more than one exposure can be assessed when general population is the source of the comparison group.
  - (a) True.
  - (b) False.
5. The standardized mortality ratio is a ratio of.
  - (a) Expected deaths in a cohort to the number of deaths observed.
  - (b) Observed deaths in a cohort to the number of deaths expected.
  - (c) Observed deaths in a cohort to the prevalence of disease.
  - (d) Observed deaths in a cohort to the incidence of disease.
6. A cohort study where outcomes have not occurred at baseline is called a.
  - (a) Retrospective cohort study.
  - (b) Prospective cohort study.
  - (c) Ambi-directional cohort study.
  - (d) Nested case-control study.
7. A retrospective cohort study is where.
  - (a) The exposures have occurred, looking forward in time.
  - (b) The exposures and outcomes have not occurred.

- (c) Only the exposures have occurred, in the past, looking back in time.  
(d) Exposures and outcomes have already occurred in the past, looking back in time.
8. The studies that are expensive to conduct and time consuming are.
- (a) Randomized control trials.  
(b) Case-control studies.  
(c) Prospective cohort studies.  
(d) Retrospective cohort studies.
9. All of the following are advantages of cohort studies, except.
- (a) Vulnerable to selection bias.  
(b) Good to examine rare exposures.  
(c) Can examine multiple outcomes are a given exposure.  
(d) Can assess causality.
10. Which of the following studies was based on a well-defined geographic area?
- (a) Nurses' Health Study.  
(b) Cardiovascular Health Study.  
(c) Atherosclerosis Risk in Communities Study.  
(d) Framingham Heart Study.
11. Choose one. "Biases can occur during the design phase and analysis phase of the study."
- (a) True.  
(b) False.
12. The best denominator to use when follow-up times differ in a cohort study is to use.
- (a) Person-time at risk as the denominator.  
(b) Total population at risk as the denominator.  
(c) Number of new cases as the denominator.  
(d) Person-time at risk as the numerator.
13. Risk ratio also known as relative risk is a measure that compares.
- (a) Risk of disease to the risk of outcome.  
(b) Risk in the exposed group to the risk in the unexposed group.  
(c) Risk in the unexposed group to the risk in the exposed group.  
(d) Risk of outcome to the risk of disease.
14. The rate ratio is based on the \_\_\_\_\_.
- (a) Rate of deaths.  
(b) Rate of exposure.  
(c) Rate of nonexposure.  
(d) Person-time contributed to the study.

15. A risk ratio of 1.0 indicates \_\_\_\_\_.

- (a) Exposure is protective.
- (b) Exposure is harmful.
- (c) Identical risk in the rates between exposed and unexposed groups.
- (d) Exposure is uncertain.

### Answer Keys

- 1. (c)
- 2. (b)
- 3. (a)
- 4. True
- 5. (b)
- 6. (b)
- 7. (d)
- 8. (c)
- 9. (a)
- 10. (d)
- 11. True
- 12. (a)
- 13. (b)
- 14. (d)
- 15. (c)

### References

1. Wang X, Kattan MW. Cohort studies: design, analysis, and reporting. *Chest*. 2020;158(1s):S72–s78.
2. Pickett J. The American heritage dictionary of the English language. 4th ed; 2001.
3. Samet JM, Muñoz A. Evolution of the cohort study. *Epidemiol Rev*. 1998;20(1):1–14.
4. Morabia A. A history of epidemiologic methods and concepts, vol. 93. Springer; 2004.
5. Hood MN. A review of cohort study design for cardiovascular nursing research. *J Cardiovasc Nurs*. 2009;24(6):E1–9.
6. Bao Y, et al. Origin, methods, and evolution of the three Nurses' health studies. *Am J Public Health*. 2016;106(9):1573–81.
7. Colditz GA, Philpott SE, Hankinson SE. The impact of the nurses' health study on population health: prevention, translation, and control. *Am J Public Health*. 2016;106(9):1540–5.
8. Celentano DD, Szklo M, Gordis L. Cohort studies. 6th ed. Gordis epidemiology; 2019.
9. Setia MS. Methodology series module 1: cohort studies. *Indian J Dermatol*. 2016;61(1):21–5.
10. Song JW, Chung KC. Observational studies: cohort and case-control studies. *Plast Reconstr Surg*. 2010;126(6):2234–42.
11. Valvi D, et al. Fibrinogen, chronic obstructive pulmonary disease (COPD) and outcomes in two United States cohorts. *Int J Chron Obstruct Pulmon Dis*. 2012;7:173–82.
12. Grimes DA, Schulz KF. Cohort studies: marching towards outcomes. *Lancet*. 2002;359(9303):341–5.
13. Szklo M. Population-based cohort studies. *Epidemiol Rev*. 1998;20(1):81–90.

14. Mahmood SS, et al. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet*. 2014;383(9921):999–1008.
15. Kannel WB. Contribution of the Framingham Study to preventive cardiology. *J Am Coll Cardiol*. 1990;15(1):206–11.
16. Meadors G. Justification for the budget estimate for the sub-project “epidemiology”. Rockville, MD/Framingham, MA: United States Public Health Service/ Framingham Heart Study Archives; 1947.
17. Chapman JM, et al. Measuring the risk of coronary heart disease in adult population groups. The clinical status of a population group in Los Angeles under observation for two to three years. *Am J Public Health Nations Health*. 1957;47(4 Pt 2):33–42.
18. Keys A. Longevity of man: relative weight and fatness in middle age. *Ann Med*. 1989;21(3):163–8.
19. Dawber TR, Meadors GF, Moore FE Jr. Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health Nations Health*. 1951;41(3):279–86.
20. Kannel W, et al. Vascular disease of the brain—epidemiologic aspects: the Framingham study. *Am J Public Health Nations Health*. 1965;55(9):1355–66.
21. Fried LP, et al. The cardiovascular health study: design and rationale. *Ann Epidemiol*. 1991;1(3):263–76.
22. Fried LP, et al. Risk factors for 5-year mortality in older adults: the cardiovascular health study. *JAMA*. 1998;279(8):585–92.
23. Muriel R-S. Limitations and biases in cohort studies. In: Barría RM, editor. *Cohort studies in health sciences*. IntechOpen: Rijeka; 2018. p. Ch. 3.
24. Aschengrau AASGR. *Essentials of epidemiology in public health*. Burlington, MA: Jones & Bartlett Learning; 2020. p. 270.

# Chapter 5

## Epidemiological Measures



Praphul Joshi

### Learning Objectives

After completing this chapter, you will be able to:

- Define epidemiology.
- Define basic measures, including incidence, prevalence, mortality, and morbidity.
- Assess crude rates and age-adjusted rates.
- Utilize the principles of epidemiological measures in describing a disease.

## 1 Introduction

Epidemiology is the study of the distribution and determinants of health-related states or events in specified populations and the application of this study to control health problems [1]. This discipline is data driven and can only be practiced by structured methods in data collection, analysis, and interpretation. To understand and implement any epidemiological study, it is very important to quantify each component of the definition of epidemiology. This chapter deals with a wide range of measures used to describe each epidemiology component. The sections below will discuss each component of epidemiology and will provide examples of different diseases for better understanding.

Being a part of life science, epidemiology involves evaluating diseases and health issues from birth to death. In this regard, the epidemiological measures range from birth to death. Health sciences aim to eradicate health issues, prevent illness/injuries, prevent death, and ultimately increase the life span. From birth to death, each level in this spectrum can be explained using epidemiological measures. This leads

---

P. Joshi (✉)

Department of Public Health, Sam Houston State University, Huntsville, TX, USA  
e-mail: [pxj015@shsu.edu](mailto:pxj015@shsu.edu)

to several measures in quantifying the spectrum of life, including fertility, birth rates, morbidity, disability, and mortality. Due to the vast diversity in the measures, specific terms in the sections below are grouped based on data sources. This will enable students and professionals to incorporate appropriate epidemiological measures using different data sources.

## 2 Crude Numbers and Rates

Each health indicator can be expressed in raw numbers (e.g., the total number of people who died due to COVID-19 in 2020 in Louisiana). However, to compare different geographic places, rates are often used for fair comparison due to differences in the overall population. For example, the population of Texas is many times higher than Mississippi. Hence, Texas is expected to have more deaths from COVID-19 compared to Mississippi. *Rates* will consider each geographic region's overall population and offer a better way to evaluate health issues. In several instances, populations of comparison may be significantly different regarding age groups and other characteristics. In this case, the crude rates need to be *adjusted* according to demographic indicators. The most common ways of adjustment include *age-adjusted* rates. If a geographic area under comparison has a predominantly geriatric population and the other area of comparison has predominantly younger age groups, the health outcomes of those two regions may be very different. Using age-adjusted rates provides a fair comparison of public health issues; examples of those rates will be discussed in the below sections as specific indicators are discussed.

Each section will provide examples of epidemiological measures using specific diseases/health issues and references for data sources, eventually enabling readers to conduct ecological studies. Basic descriptive statistics and online tools to enable data gathering and interpretation of data will be described.

### 2.1 *Rates, Ratios, and Proportions*

A rate is the measure of the frequency of an event for a specific population during a period. Rates are used in a variety of settings, and most epidemiological measures are expressed in rates. Examples of rates include fertility, mortality, case fatality, incidence, and prevalence rates. Some of the rates will be explained in this chapter, while others will be explained in other chapters depending on the epidemiological context.

A ratio is the relative magnitude of two quantities or a comparison of any two values. It is calculated by dividing one interval- or ratio-scale variable by the other. It helps to compare populations/subgroups, including age, gender, race/ethnicity, income levels, or educational attainment.

$$\text{Ratio} = \frac{\text{Number of events in a population}}{\text{Number of events in a comparison population}}$$

A proportion compares a specific population subgroup to the overall population. It can be used in many combinations of demographic factors and disease indicators. For instance, the rate of lung cancer among African Americans compared to that of the rest of the population. In this case, the proportion can be calculated as (Box 5.1)

$$\text{Proportion} = \frac{\text{Rate of lung cancer among African Americans}}{\text{Rate of lung cancer among overall population}}$$

#### **Box 5.1: COVID-19 Examples for Rates, Ratios, and Proportions**

Ratio is a common measure used in descriptive epidemiology. Examples of ratios for epidemiological studies include risk ratio, odds ratio, and death-to-case ratio. Death-to-case ratio has been used in COVID-19 to determine the number of deaths in a community due to the disease divided by the total number of new cases with a confirmed disease during the same period.

Rates are used in numerous contexts in epidemiology including incidence rate, prevalence rate, vaccination rate, and case-fatality rate. In COVID-19, vaccination rates were frequently used to determine the outcomes (prevention of acute hospitalization and deaths). In this case, the vaccination rate is calculated using total number of people receiving the vaccine divided by the eligible population.

Proportions have been extensively used in case of COVID-19. For instance, the rate of mortality due to COVID-19 among those who are obese compared to the mortality due to COVID-19 in rest of the population.

### **3 Measures Used in Vital Statistics**

The most common source of data to drive public health efforts is vital statistics, which includes data from birth and death records. Every public health department and county in the United States collects vital statistics which are reported to their respective state health departments and ultimately to the Centers for Disease Control and Prevention (CDC). The Office of the National Center for Health Statistics under the CDC is the go-to place for evaluating data from vital statistics (NCHS). Common epidemiological measures used from vital statistics and some examples will be discussed below.

### 3.1 Birth Data

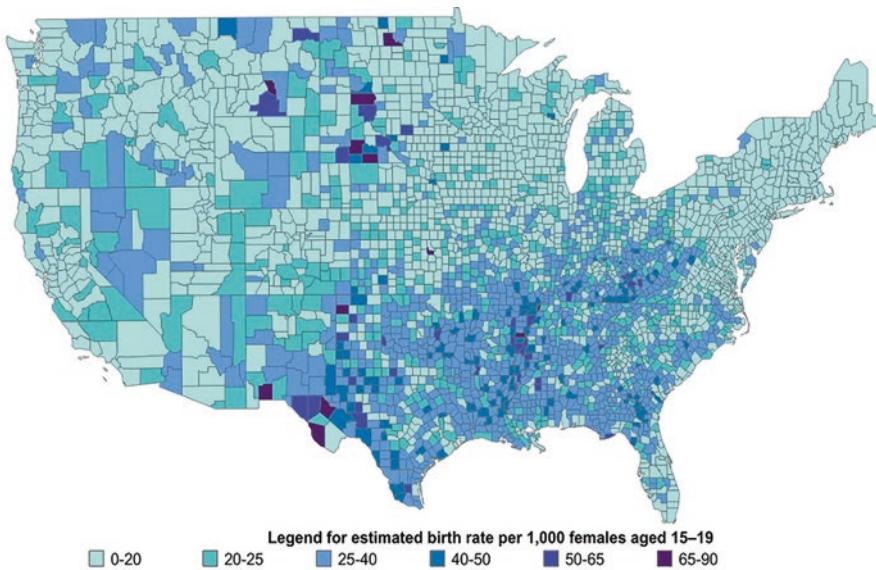
Each person has a birth certificate, and every birth includes a large amount of data that is extremely helpful in understanding a wide range of public health issues. Each birth adds to the overall population, and the number of births per year is the primary driver for the population of every community. Typically, the overall birth for any given community is quantified as the birth rate. The birth rate is defined as the total number of births in a year divided by the total population. For larger population entities (large cities, states, or countries), the rate is often expressed as the number of births per 1000 population. This indicator is also referred to as the *crude birth rate*, as it is calculated using the overall population as the denominator. This indicator can also be expressed as an *age-adjusted birth rate* if specific age groups within the population are considered, which offers a fair comparison across two geographic regions. The number of births is also used to define another term, *fertility rate*, which is the total number of births divided by the number of women in reproductive each (15–50 years). The total fertility rate indicates the number of children a woman would have had in the reproductive years (typically between 15 and 50 years).

A critical component of the birth certificate is the age of the mother (maternal age). This is used in calculating the rates of teen pregnancies. *Teen birth rates* are the total number of live births born to women ages 15–19 divided by the population of women in that age group for a particular year. Like crude birth rates, teen birth rates are also typically calculated per 1000 females aged 15–19 years. Figure 5.1 describes the teen birth rates in the counties in the United States for the year 2020. Figure 5.2 describes the teen birth rates by state for the same year (2020).

Birth data also captures several other characteristics of the mother, including demographic factors (age, race/ethnicity, and geographic area), underlying health issues (e.g., gestational diabetes), kind of delivery (natural or surgical intervention), place of delivery (e.g., home or hospital), socioeconomic indicators (income, health insurance status, and education), and marital status. All these factors can be utilized in evaluating the birth data based on maternal characteristics. An example of trends in births to unmarried women in the United States is presented in Fig. 5.3.

### 3.2 Mortality Data

Like birth data, mortality data is obtained from death records and is reported to various levels from county, state, and nationwide. The mortality rate is defined as the total number of deaths each year divided by the overall population. This indicator is also expressed as the rate for 1000 population. Death records indicate many variables critical in implementing epidemiological studies and making public health decisions. Age is the most significant contributor to mortality; hence, communities with aging populations are expected to have higher mortality rates. To enable fair comparisons across communities, *age-adjusted mortality rates* are often used to

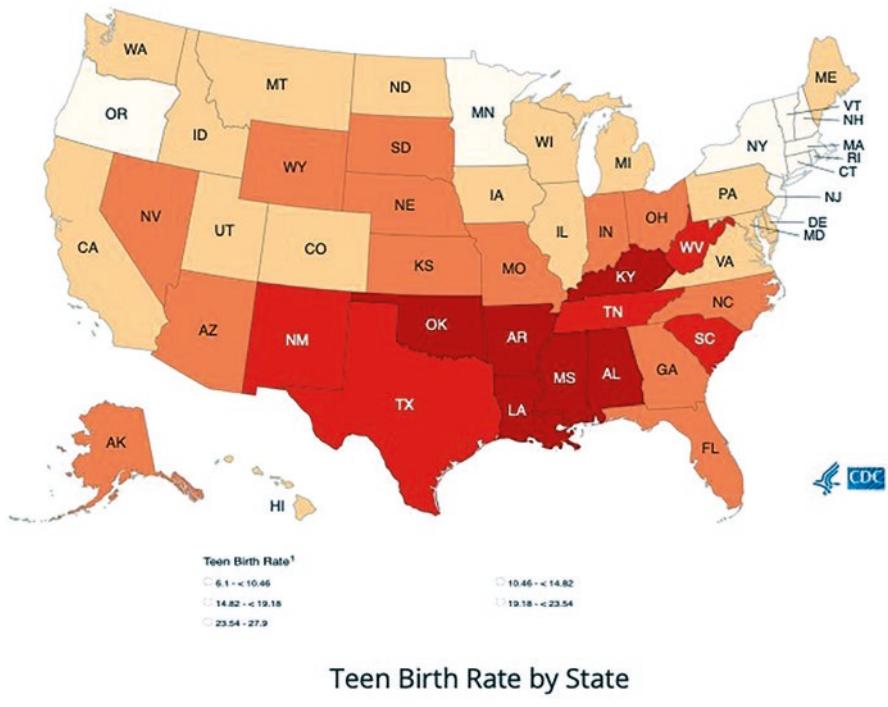


Estimated Teen Birth Rates for Females Aged 15–19 by County: Continental U.S., 2020

**Fig. 5.1** Estimated teen birth rates for females aged 15–19 years by county in the United States for 2020. (Source: The National Center for Health Statistics, CDC [3])

describe most diseases and public health issues. Figure 5.4 describes the differences between the overall number of deaths, crude mortality rates, and age-adjusted mortality rates. As seen, the trends appear very different based on each indicator used. Interpretation of these terms is very critical in disseminating public health messages. The overall number of deaths is cumulative and is only expected to increase. Demographic changes in population (migration, age changes), public health measures, and emerging health issues play a major role in the trends of mortality rates. As indicated in the figure, the age-adjusted mortality rate decreased at a greater rate than the crude rates. This was due to an overall increase in life expectancy in the United States.

Mortality age is a primary factor in determining life expectancy, defined as the number of years a person is likely to live. Mortality can happen at any age from birth. Any additional year past birth contributes toward life expectancy. Typically, the higher a community's life expectancy, the better the public health infrastructure. On the contrary, poor public health infrastructure leads to lower life expectancy, and many might die at birth (Infant mortality). The *infant mortality rate* is calculated as the number of infant deaths for every 1000 live births. Figure 5.5 represents the infant mortality rate in the United States for the year 2020 across the states. Other related terms used in public health include *neonatal mortality rate* (number of deaths from birth through 27 days of life), the *post-neonatal mortality rate* (number of deaths between 28 and 364 days), and *perinatal mortality rates* (number of deaths of fetus older than 28 weeks of gestation till seven days of birth). Each of these

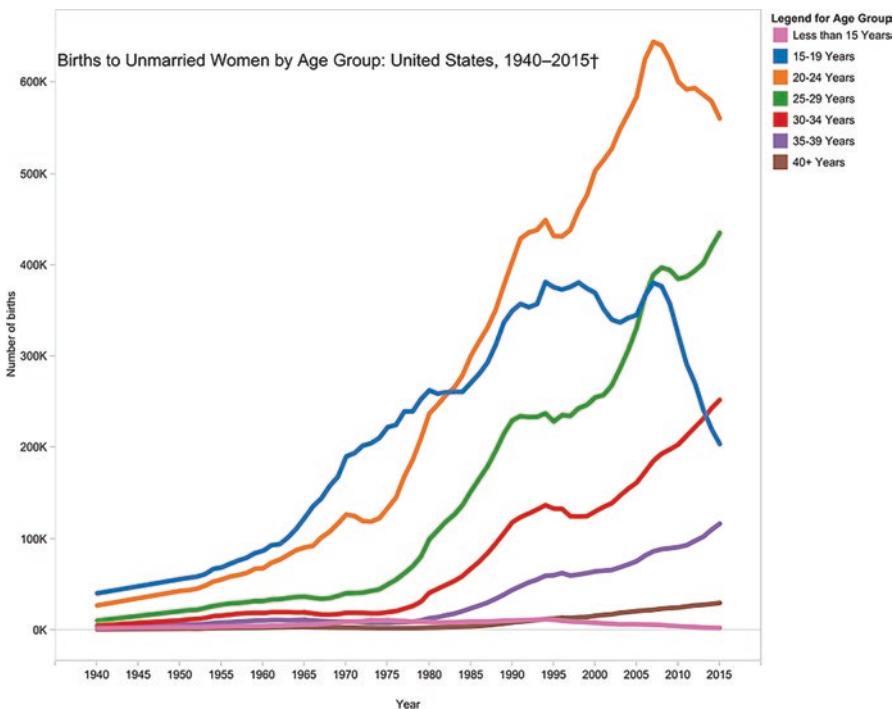


**Fig. 5.2** Teen births by state in the United States for 2020. (Source: The National Center for Health Statistics, CDC [3])

terms has specific public health implications and should enable in the data-driven decision-making process.

*Cause-specific mortality rates* indicate the number of deaths due to a particular disease or a health issue per 1000 population each year. All deaths must be certified by a physician in the United States and reported to the local authorities. Like birth records, death records contain a wide range of information that can be used to evaluate public health issues. As indicated above, age at mortality is a critical factor that determines the life expectancy of a population. Death records include the primary cause of death, typically coded based on the International Classification of Diseases (ICD Codes). ICD-10 is the latest classification of disease code used to determine primary causes of death. A large proportion of deaths also have a secondary cause of death, which also will be appropriately coded based on ICD guidelines. Other information included in the death records includes place of death (city, house/hospital), date of death (sometimes may have to be estimated), demographic information (gender, race/ethnicity), socioeconomic indicators (e.g., occupation, income, and health coverage), marital status, and other information depending on causes (e.g., type of injury and accident).

Figure 5.6 describes the cause-specific mortality rate due to chronic obstructive pulmonary disease (COPD) in the United States from 2001 to 2020. The crude rates



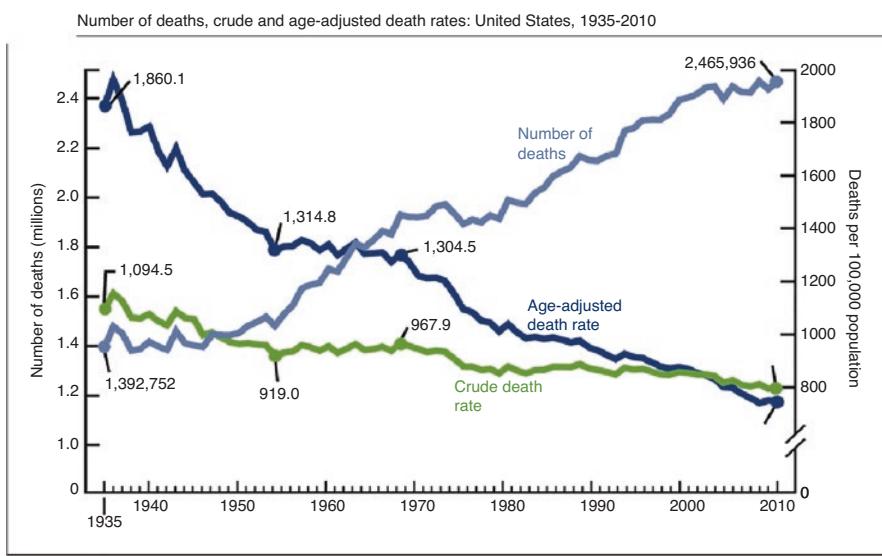
**Fig. 5.3** Number of births to unmarried women by age group in United States from 1940 to 2015. (Source: Ventura and Bachrach, 2000 [2])

appear to be increasing over time, whereas the age-adjusted rates appear to have no change or slight decline.

The differences between crude and adjusted rates for COPD can be explained if the mortality rates among specific age groups over the past 20 years were to be analyzed. Figure 5.7 describes the trends in COPD mortality rates based on 10-year age groups from 2001 to 2020. It should be noted that the most significant declines in mortality have been observed in recent years among the older age groups (75–84 years and 85+ years). This explains the changing trends in age-adjusted rates, as reported in Fig. 5.6 (Box 5.2).

### Box 5.2: Health Indicators

Health indicators describe the public health status of a community. They are like diagnostic tools and test results in a clinical setting. Health indicators help in diagnosing specific public health issues and provide ways to implement appropriate interventions. Health indicators help in formulating epidemiological hypothesis, enable utilizing appropriate study designs, and enable in improving the overall health of the community.



NOTES: 2010 data are preliminary. Crude death rates on an annual basis are per 100,000 population; age-adjusted rates are per 100,000 U.S. standard population. Rates for 2001–2009 are revised and may differ from rates previously published.

SOURCE: CDC/NCHS, National Vital Statistics System, Mortality.

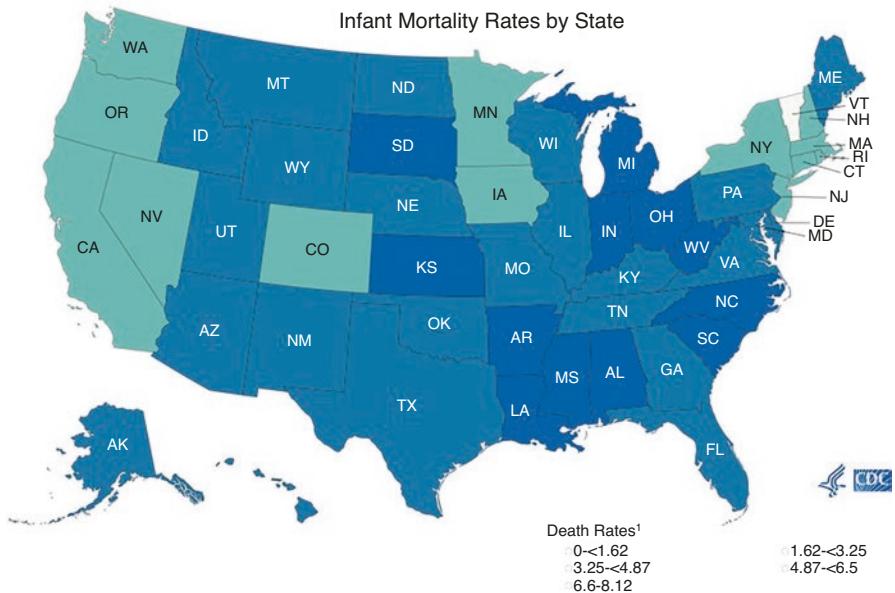
**Fig. 5.4** Comparison of overall deaths, crude mortality rates, and age-adjusted mortality rates in the United States from 1935 to 2010. (Source: Centers for Disease Control and Prevention [3])

Significant advancements in treatment options, preventative screenings, and increased public health education efforts are expected to significantly decrease cause-specific mortality rates. A classic example is the decrease in lung cancer mortality rates. Over the past three decades, immense public health education efforts have led to significant decreases in smoking behaviors and, subsequently, in lung cancers. Medical advancements and enhanced screenings have also contributed to the prevention of death from lung cancers over the past two decades.

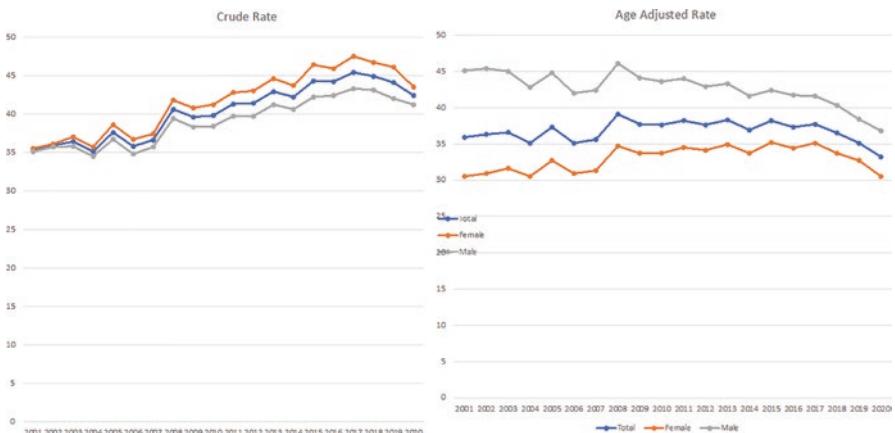
## 4 Morbidity Measures

Morbidity includes the states of altered physical or psychological well-being that can lead to disease, disability, or impact on quality of life. Measures of morbidity include all spectrum of variables between birth till death. The following sections will discuss several epidemiological measures used to describe the states of morbidity.

*Prevalence* represents the total number of current cases having a specific disease. The number of people with a particular disease divided by the population will provide the *prevalence rate*. As with other rates discussed in previous sections, prevalence rates are typically calculated per 1000 population. Two other terms are used to describe the prevalence of a disease. *Point prevalence* refers to the number of cases of a particular disease at a particular point in time. Period prevalence represents the



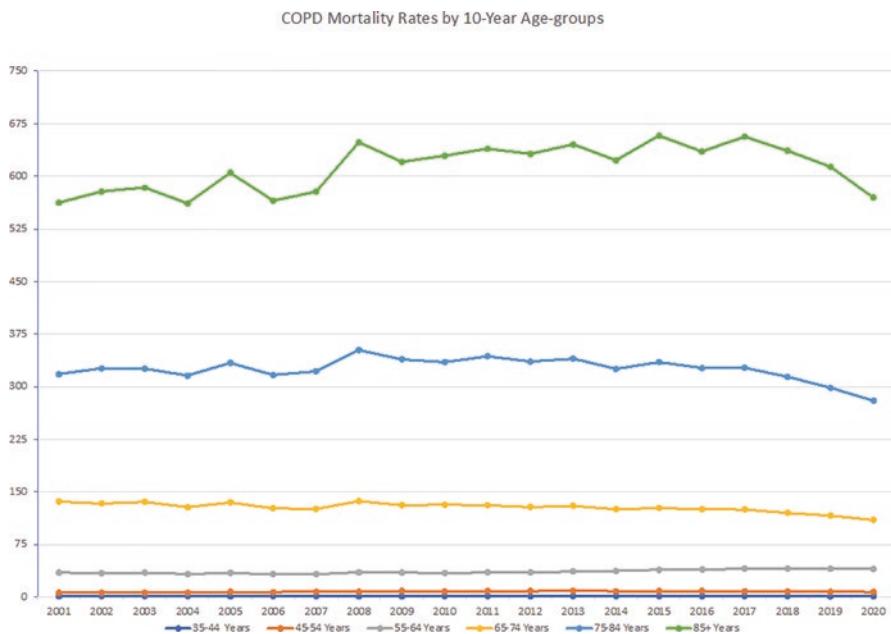
**Fig. 5.5** Infant mortality rates in the United States across the states for the year 2020. (Source: National Center for Health Statistics, CDC (2000) [3])



**Fig. 5.6** Differences between crude and adjusted cause-specific mortality rates due to chronic obstructive pulmonary disease (COPD) in the United States from 2001 to 2020. (Data Source: CDC Wonder ([wonder.cdc.gov](http://wonder.cdc.gov)) [4])

number of cases (old and new) of a particular disease in a period (week, month, or year).

*Incidence* refers to the number of new cases of a particular disease in a given population. *Incidence rate* is calculated by using the number of new cases of a particular disease in a period divided by the overall population. These rates are also



**Fig. 5.7** Trends in COPD mortality rates by 10-year age groups from 2001 to 2020. (Data source: CDC Wonder ([wonder.cdc.gov](http://wonder.cdc.gov)) [4])

typically calculated for 1000 population. The period for incidence rates may be weekly or monthly for emerging infectious diseases (COVID-19) or may be calculated annually for other diseases (cancers). *The attack rate* is calculated by using the number of new cases divided by the population's size at the beginning of the period. Both incidence and attack rates indicate the severity of the disease and how fast a disease can spread in each population (Box 5.3).

### Box 5.3: Incidence and Prevalence

From the above sections, it can be summarized that prevalence is incidence rate times the duration of the disease.

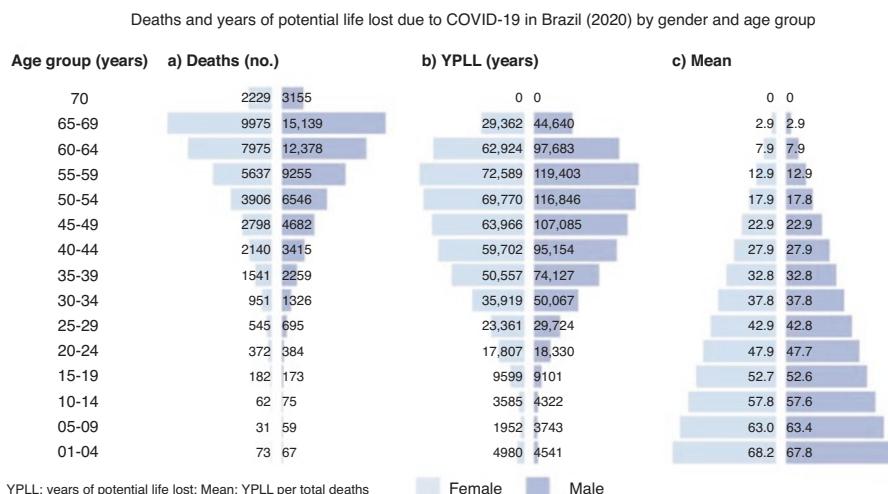
$$\text{Prevalence} = \text{Incidence} \times \text{Average duration of a disease}$$

Using the above equation, average duration of the disease can also be calculated using incidence and prevalence rates:

$$\text{Average duration of disease} = \frac{\text{Prevalence}}{\text{Incidence}}$$

For chronic diseases (hypertension, diabetes, or cancers), the number of new cases keeps adding each year. People with most chronic diseases live with them (rather than getting fully cured); hence, chronic diseases' prevalence increases over time. For infectious diseases with high mortality rates (e.g., Ebola) or very fast recovery (e.g., flu), disease prevalence can decrease when in control and can increase during outbreaks. Figure 5.8 represents the incidence, prevalence, and mortality due to HIV and those with antiretroviral therapy.

The prevalence of any disease increases morbidity and can lead to disability, death, or both. Improvement in public health measures not only decreases mortality rates but also aims to reduce disease burden. Years of potential life lost (YPLL) is a good indicator to measure the disease's burden. This indicates the average number of years a person would have lived if they did not die from a particular disease. Premature death is often-times considered as death before 65 years of age. Typically, a person's age at death is subtracted from 64.5 years. For example, if a person dies at the age of 50, the YPLL will be  $64.5 - 50 = 14.5$  years. The YPLL for a particular disease can be averaged for individuals in each state or a geographic region to calculate YPLL for a particular disease in each state or country. Figure 5.8 indicates the YPLL due to COVID-19 in Brazil. Younger the age of death, the greater will be YPLL – the older the age of death, the lower will be YPLL. Mirror images of the number of deaths per age group and YPLL per age group are represented in the figure.



**Fig. 5.8** Years of potential life lost (YPLL) due to Covid-19 by gender in Brazil in 2020.(Source: Castro et al. 2021 [5])

## 5 Useful Online Resources

CDC Wonder: [wonder.cdc.gov](http://wonder.cdc.gov). Wide-ranging Online Data for Epidemiologic Research – an easy-to-use, menu-driven system that makes the information resources of the Centers for Disease Control and Prevention (CDC) available to public health professionals and the public at large. It provides access to a wide array of public health information.

Behavioral Risk Factor Surveillance System (BRFSS): [www.cdc.gov/brfss](http://www.cdc.gov/brfss). Nation's premier system of health-related telephone surveys that collect state data about U.S. residents regarding their health-related risk behaviors, chronic health conditions, and use of preventive services. Established in 1984 with 15 states, BRFSS now collects data in all 50 states, the District of Columbia, and three U.S. territories. BRFSS completes more than 400,000 adult interviews each year, making it the world's largest continuously conducted health survey system.

Youth Risk Behavior Surveillance System (YRBSS): [www.cdc.gov/yrbs](http://www.cdc.gov/yrbs). The YRBSS monitors health behaviors that contribute markedly to the leading causes of death, disability, and social problems among youth and adults in the United States. These behaviors are often established during childhood and early adolescence.

National Center for Health Statistics: [www.cdc.gov/nchs](http://www.cdc.gov/nchs). NCHS maintains online databases for vital statistics, population-based surveys, and provider-based surveys and hosts interactive tools to evaluate a wide range of public health issues.

## 6 Further Practice: Case Study

1. Visit the website [wonder.cdc.gov](http://wonder.cdc.gov) and obtain the crude death rates for breast cancer for a state of your choice for the years 2011–2020. Compare this data with the age-adjusted rates for the same years. Explain the differences in your findings.
2. Visit [www.cdc.gov/brfss](http://www.cdc.gov/brfss) and obtain the prevalence of smoking among the adult population in the state of your choice over the last decade. Compare this trend with the rest of the United States. Explain your findings.
3. Brianna is investigating lung cancer in Maricopa County. The county has 240,000 overall population. She visits the local cancer registry and notices that there are 1200 people in the county with the disease. In the year 2021, 120 people in the county died due to lung cancer, and in the first six months of 2022, 40 people died due to lung cancer. Over the last five years, there have been about 310 new lung cancer cases. Answer the following questions:
  - (a) Calculate is the prevalence rate of lung cancer in Maricopa County in 2022.
  - (b) Calculate is the disease-specific mortality rate due to lung cancer in 2021.
  - (c) What is the average incidence of lung cancer each year in Maricopa County?

**Multiple Choice Questions**

1. The prevalence of a disease increases when.
  - (a) Mortality is high.
  - (b) Recovery time is fast.
  - (c) Rapidly spreading disease agent.
  - (d) Low spreading disease agent.
2. A COVID-19 outbreak can increase.
  - (a) Incidence rate.
  - (b) Prevalence rate.
  - (c) Cumulative mortality.
  - (d) All the above.
3. Amy is studying trends in the prevalence of prostate cancer by race. Health disparities exist when:
  - (a) The lines are intersecting.
  - (b) Increasing parallel lines.
  - (c) Steady parallel lines.
  - (d) All the above.
4. What is the denominator for fertility rate.
  - (a) Overall number of births in a year.
  - (b) Overall population in the county.
  - (c) Number of women in the county.
  - (d) Population of women between 15 and 49 years.
5. What is the numerator for birth rate.
  - (a) Overall population of the county.
  - (b) Number of women in the county.
  - (c) Total number of births in the county.
  - (d) Total number of pregnancies in the county.
6. YPLL increases with:
  - (a) Increase in life expectancy.
  - (b) Mortality in younger age groups.
  - (c) Fewer deaths in elderly.
  - (d) All the above.
7. Which of the following diseases would have the highest YPLL:
  - (a) Alzheimer's disease.
  - (b) Parkinson's disease.
  - (c) Fatal road traffic accidents.
  - (d) Brain cancer.

8. A total of 1500 new cases of COVID-19 were reported in Jackson County in the month of October 2021. This represents:
  - (a) Incidence.
  - (b) Prevalence.
  - (c) Mortality.
  - (d) YPLL.
9. As of October 31, 2021, 85% of the residents of Richmond County had received at least two doses of COVID-19 vaccine. This represents:
  - (a) Incidence.
  - (b) Prevalence.
  - (c) Mortality.
  - (d) YPLL.
10. COVID-19 has led to five fewer years of life for people aged less than 65 years in Polk County. This represents.
  - (a) Incidence.
  - (b) Prevalence.
  - (c) Mortality.
  - (d) YPLL.

### Answer Keys

1. (c)
2. (d)
3. (d)
4. (d)
5. (c)
6. (d)
7. (c)
8. (a)
9. (b)
10. (d)

### References

1. Last JM, editor. Dictionary of epidemiology. 4th ed. New York: Oxford University Press; 2001. p. 61.
2. Ventura SJ, Bachrach CA. Nonmarital childbearing in the United States, 1940–99. Natl Vital Stat Rep. 2000;48(16):1–40. Available at: [https://www.cdc.gov/nchs/data/nvsr/nvsr48/nvs48\\_16.pdf](https://www.cdc.gov/nchs/data/nvsr/nvsr48/nvs48_16.pdf)
3. National Center for Health Statistics. Infant mortality rate. Available at: [https://www.cdc.gov/nchs/pressroom/sosmap/infant\\_mortality\\_rates/infant\\_mortality.htm](https://www.cdc.gov/nchs/pressroom/sosmap/infant_mortality_rates/infant_mortality.htm)
4. CDC Wonder. Available at: <https://wonder.cdc.gov/>
5. Castro APB, Moreira MF, Bermejo PHS, Rodrigues W, Prata DN. Mortality and Years of Potential Life Lost due to COVID-19 in Brazil. Int J Environ Res Public Health. 2021;18(14):7626. <https://doi.org/10.3390/ijerph18147626>.

# Chapter 6

## Clinical Trials



Michael Bennish and Wasif Ali Khan

### Learning Objectives

After completing this chapter, you will be able to:

- Understand the key elements of a clinical trial
- Be familiar with the ethical concerns in human subject research.
- Understand the role of the sponsor of clinical trials
- Know the concept of equipoise in determining when planning a clinical trial

## 1 Introduction

Historically, health interventions were often based upon commonly accepted practices that seemed intuitively to make sense and perhaps had some biological plausibility. However, evidence that they improved health was often limited or nonexistent. The examples of such practices are myriad. The use for many centuries of bleeding to “treat” a variety of ailments, or in my own lifetime, the routine administration of tonsillectomies and adenoidectomies to all children, are only two examples of common medical practices that have been in due course shown not to confer benefit, while posing risks to patients. Combined with the pecuniary interest that medical practitioners have traditionally had in providing interventions, it is not surprising that the playwright and essayist Bernard Shaw famously (and acerbically) wrote more than a hundred years ago in *The Doctor’s Dilemma* [1] that

---

M. Bennish  
Mpilonhle, Mtubatuba, South Africa  
e-mail: [michael@mpilonhle.org](mailto:michael@mpilonhle.org)

W. A. Khan (✉)  
icddr,b, Dhaka, Bangladesh  
e-mail: [wakhan@icddrb.org](mailto:wakhan@icddrb.org)

“...the rank and file of doctors are no more scientific than their tailors...” and

“That any sane nation, having observed that you could provide for the supply of bread by giving bakers a pecuniary interest in baking for you, should go on to give a surgeon a pecuniary interest in cutting off your leg, is enough to make one despair of political humanity.”

Medicine has come a long way in the past hundred years, and the emphasis on evidence-based medicine and the increasing use of artificial intelligence in clinical decision-making will only increase the need for reliable evidence to use when choosing health interventions. Clinical trials are the gold standard for providing such evidence and are the focus of this chapter.

International and national regulatory agencies responsible for licensing pharmaceuticals established the International Conference (now Council) for Harmonization (ICH) with a specific mandate to establish common standards for clinical research [2]. The ICH developed Good Clinical Practice Guidelines “to provide an international ethical and scientific quality standard for designing, conducting, recording and reporting trials that involve the participation of human subjects.” Importantly, “Compliance with this standard provides public assurance that the rights, safety, and well-being of trial subjects are protected, consistent with the principles that have their origin in the Declaration of Helsinki” [3]. Although ICH was established initially by the European Union and the United States, its guidelines have been widely adopted with minor variations in details and emphasis [4].

The literature on clinical trial design, conduct, analysis, and reporting is voluminous. This chapter will highlight crucial aspects of clinical trials relevant to clinicians and nonclinicians, including those who use clinical trials in their epidemiologic work.

## 2 What Constitutes a Clinical Trial?

For the first time, the 1962 Drug Efficacy Amendment to the United States Federal Food, Drug, and Cosmetic Act required that drug manufacturers demonstrate both the safety and efficacy of the drugs. They wanted to market and began to set standards for clinical trials used in support of new drug applications to the United States Federal Drug Administration (FDA) [5]. This act was also the origin of the clinical studies’ phase 1–4 categorization (Table 6.1).

As part of that effort, the NIH (a major funder of clinical trial research) and FDA, for regulatory purposes, adopted a common definition of what constitutes a clinical trial. The current definition is as follows:

“A research study in which one or more human subjects are prospectively assigned to one or more interventions (which may include placebo or other control) to evaluate the effects of those interventions on health-related biomedical or behavioral outcomes.” [6]

**Box 1**

Of note, a clinical study, as opposed to a limiting definition of a clinical trial, does not require a comparison group. Some of the most fundamental therapies – insulin in type 1 diabetes or penicillin for treating wound infections or pneumococcal pneumonia – were never subject to controlled trials [7, 8]. The initial prospective and carefully observed studies of insulin and penicillin were compared to historical experience, and their dramatic effects were obvious.

**Table 6.1** Phases of a clinical trial of a new drug, biological process, or device

Study phase	Purpose	Usual no. of participants
Phase 1	To evaluate the safety of the drug, biological product, or device	10–50
Phase 2	Preliminary testing of efficacy and dosing amount and more detailed safety information. Usually, no comparison intervention is evaluated and not powered to detect a clinically significant effect of the intervention	<100
Phase 3	Comparative study of the efficacy of intervention; compare with standard treatment (or placebo). Important that has sample size is large enough that the study has sufficient power to show the potential benefit of the intervention	100–1000 s
Phase 4	A long-term study of side effects and benefits after licensing of a drug, biologic product, or device	>1000

There is an argument that clinical trials have become an obsession and are conducted even when the results, according to some, should be obvious. This has been highlighted by “spoof” articles on the randomized controlled studies of parachutes versus no parachutes when jumping from a plane. This argument has been debated in clinical medicine and perhaps even more intensely in development economics [9]. Dramatic interventions in which the results are immediately apparent, such as using penicillin to treat infections, are now rare. After more than 80 years of use, penicillin is still subject to clinical trials to test its efficacy for conditions where the outcome is less certain and dramatic than when first used [10]. Thus, randomized, controlled trials (RCTs), rather than personal experience and uncontrolled observations, should now serve as the basis for clinical decision making.

### 3 Study Design

#### 3.1 Superiority, Equivalence, and Noninferiority Trials

When developing the hypothesis for a study, it is important to define what is to be “proved” by the study. This is fundamental for making the study results interpretable to the community that might use or benefit from the results. It is also critical for

determining the study's sample size and for the study's results to be considered valid and generalizable and thus of use to practitioners and public health officials for making informed decisions.

The majority of investigators hypothesize that a new intervention is superior to current practice or treatment. For example, a new antihypertensive agent provides greater benefits than the current therapy. Either because it results in a greater (and clinically relevant) decrease in blood pressure or is associated with fewer side effects or toxicity). Because an underlying assumption for clinical trials is the state of equipoise exists – i.e., there is uncertainty as to any possible benefit of the tested intervention in relation to the current practice (see below) – “superiority” studies are also “inferiority” studies. If equipoise truly exists, there is an equivalent probability that the test intervention is worse than current practice. For this reason, all tests of the significance of the study outcome must be two tailed. It means – if the assumption is that there is a 95% probability that the differences between interventions did not occur by chance alone – that both sides of the tail of the normal distribution have been tested to see if the results are in the top 2.5% or the bottom 2.5% of the normal distribution.

A noninferiority trial aims to show that the intervention being examined in the clinical trial does not perform less well than the comparator, standard intervention. As such, the test of the significance of differences between groups is usually one sided, thus effectively reducing the sample size needed by half.

An equivalence trial aims to show no important difference between the investigational and comparator interventions. In most studies, the comparison intervention will be the current standard, practice, or placebo if there is no currently identified effective intervention. The important context is that there are “no important” (or clinically significant) differences. A difference of  $\pm 1$  mm Hg is unlikely to be considered clinically significant for blood pressure reduction. Thus, an equivalence study may be designed to show the effect on blood pressure of the intervention, and comparison treatments have a probability of 95% (assuming that a  $p < 0.05$  is utilized) of differing by no more than  $\pm 1$  mm Hg. With a large enough sample size, there is likely to be a “statistically significant” but “clinically insignificant” difference between any two interventions that are compared, as few are likely to have the same effect.

### **3.2 Parallel Group, Crossover, and Factorial Study Designs**

*Parallel group* designs are the most common and the simplest study design. In a parallel group design, study subjects are assigned to different interventions.

In a *crossover study*, participants receive the study interventions sequentially. Crossover studies that examine clinical response require a washout period, in which any effect of the initial treatment is “washed out” so that the second treatment provided during the crossover study can be examined independently. In practice, this is hard to accomplish in studies that examine efficacy, as any benefits from the first

intervention are difficult to disentangle from the effect of the second treatment, such as if the comparative effects of diet and pharmacotherapy were compared for their effect on hypertension. Studying the pharmacokinetics of drugs is more amenable to a crossover study design. There are objective measures that the first drug has been “washed out” and thus would have little effect on the pharmacokinetics of the second drug being studied. In a crossover study, the advantage is that each study subject can serve as its control.

*Factorial designs* study a number of different interventions in varying combinations. For instance, a  $2 \times 2$  factorial design of weight loss and intensive pharmacotherapy for moderate hypertension might include four groups who would receive a weight loss intervention alone; intensive pharmacotherapy alone; both interventions simultaneously; and standard therapy. .

## 4 Sample Size Determination

Methods for calculating sample size are discussed in detail in Chap. 17. It is important, however, to incorporate a number of consistent principles for determining sample size. There has to be a best estimate of the event rate under study in the population sampled. If there is no relevant literature, study investigators should survey the population to obtain a reliable estimate of the incidence or prevalence of the condition of interest. The difference between the study intervention and the control group (or standard intervention group) must be clinically relevant. With a large enough sample size, it is possible to show a difference between most interventions. The difference found, however, may not have any import for practice. This is not to imply that equivalence studies have no value – but the import of the equivalent intervention (a less costly intervention, for instance) must justify the clinical trial.

The study sample size must be large enough to determine if a true difference exists between interventions. This is referred to as the “power” of the study and is usually set at 80% or 90%. A study with a power of 80% has an 8 in 10 chance of finding a difference between study groups if a difference actually exists. Failure to find a difference when a true difference exists (falsely accepting the null hypothesis) is termed a “Type II” error. A “Type I error” falsely rejects the null hypotheses. The higher the power and the lower the *P* value, the larger the sample size.

Lastly, sample size determination must consider the number of study subjects who are not likely to reach the endpoint. Even if an intention-to-treat analysis is used, there need to be enough study subjects who reach the study endpoint for the proposed analysis to be valid. The higher the dropout or noncompliance rate, the larger the sample size. The allocation ratio (one-to-one, two-to-one, etc.) also will affect sample size.

## 5 Essential Elements of a Randomized Clinical Trial

Standardizing clinical trial design has allowed for consistent interpretation and use of trial results and comparison between trials. This is particularly true for drug and device trials, whose approval is based upon having standardized protocols and study design.

### 5.1 Standardized Study Protocol and Registration of a Study

There are a number of sets of guidelines for the design of RCTs issued by various regulatory authorities and international organizations [11]. All RCTs should use a standardized protocol that is submitted to a study trial registry in advance of the initiation of the study. The largest primary international registry – [www.clinicaltrials.gov](http://www.clinicaltrials.gov) – is maintained by the United States National Library of Medicine and, at the time of writing, has more than 400,000 clinical studies from 221 countries registered with it [12]. There are other important national and regional registries, including for the European Union [13] and for Africa [14]. There is also the WHO-maintained site that collates data from a number of different registries [15]. Many countries require that clinical studies be registered with their national directory.

Study registries serve a number of important purposes. They help ensure that in the analysis of study outcomes, researchers adhere to a priori hypotheses rather than conducting post hoc analyses that are subject to researcher bias. They also allow national governments to know what research is being conducted within their borders. This is an especially important consideration in lower income countries, where much of the research is funded by external sources, and the suspicion is that local residents are being exploited for the benefit of others. Registries allow investigators contemplating a study to know what clinical studies are underway or completed but are yet to be published. Thus, avoiding duplication of effort.

### 5.2 Hypothesis and Outcome Measures

It is essential to have a well-defined a priori hypothesis and outcome measures for testing in the clinical trial. For instance, if the hypothesis is that the intervention being tested (a drug, weight loss, exercise) will reduce blood pressure, the outcome measure has to be specific. The hypothesis needs to state the amount that either the diastolic or systolic blood pressure will fall; after how long an intervention this will be measured; how sustained the fall in blood pressure will be; and how many blood pressure measurements will be used to determine the endpoint of a decrease in blood pressure. This specificity is crucial for assuring that investigator bias does not

affect the reporting of the study results. The investigation chooses an endpoint after study completion that is most conducive to showing the intervention in the best possible light.

### ***5.3 Equipoise Between the Intervention and Comparison Group***

Clinical equipoise exists when there is no definitive evidence to support the superiority of the intervention in a clinical trial over the comparator group. Based on existing information, the null hypothesis exists (i.e., the interventions compared are equivalent in their efficacy and safety), and the study intends to disprove the null hypothesis. The existence of equipoise is an essential element for the ethical conduct of clinical trials. If there is clear evidence that one of the intervention arms was superior, it is unethical to conduct the clinical trial, as one group would knowingly be receiving inferior therapy.

Determining if equipoise exists is not straightforward. One consideration is how much emphasis to place on previous studies, especially if there is only a single previous study of the intervention under consideration. There is ample evidence of considerable variance between studies. Thus, the results of a previous study may not be determinative.

The impetus for most studies is the assumption that the intervention being evaluated will be superior to the comparator standard therapy. The study results will reject the null hypothesis. In addition, it is unlikely that there is “personal equipoise” among the investigators conducting the study. As they are often involved in inventing or designing the intervention to be tested and thus are motivated by the belief that it is superior to currently used therapies. Belief does not equal evidence, however, and thus the need for clinical trials if there is insufficient evidence, after a thorough review of the published literature and accessible unpublished literature, that the proposed therapy is indeed superior in efficacy to current therapy.

A related problem is the conduct of studies in poor, resource-constrained communities where the standard of care differs from that in rich countries. It is ethical to compare a new intervention to the current standard of care in the community where the study is being conducted [16]. Should the comparison be to the higher standard available in more affluent communities? Even though it is unlikely that standard will ever be available in the community where the study is being conducted. There are increasing arguments that communities where the study will be conducted, should have a role in what studies are conducted, and how they are conducted [17].

## ***5.4 Blinding (Masking) of Study Interventions***

Blinding (or masking) refers to keeping persons unaware of which study interventions are provided to study subjects. A single-blinded study refers to assuring that the study subject herself of himself being unaware of which of the study intervention they are receiving. A double-blinded study refers to the study subject and study personnel being unaware of the intervention provided.

Analysis of study outcomes should be conducted blinded to the intervention provided to the groups compared in the analysis. This requires someone not involved with study implementation, analysis of results, unblind the study, and group participants by the intervention they received.

Blinding is important to reduce bias – intended or subconscious – in determining study outcome. Investigators for the most part are advocates of the experimental intervention that they are studying. They often have developed, studied, or promoted one of the interventions under study. Most, but not all, studies have shown a larger treatment effect in nonblinded than in blinded studies [18, 19].

Effective blinding, especially double blinding, is easier conceptually than in practice. Blinding is presumably easiest in drug intervention trials. It is hard, for instance, to blind surgical versus nonsurgical interventions or different behavioral interventions. Even in drug trials, drugs being compared may differ in taste, smell, consistency, and side effects. This may, in part, be overcome by using a double-dummy technique. Each treatment group is given one active agent and one inactive agent to resemble the other agent used in the study. Even this will not overcome the potential for side effects – such as diarrhea – that are more common with one agent – to bias investigators involved in the study.

Lack of blinding is not an absolute impediment to the validity of study results. To the degree possible, outcomes should be clearly defined with a reliably measurable outcome. In a study that compared weight-loss intervention and drug therapy to reduce blood pressure, even though the intervention cannot be blinded from either the study participant or investigative staff. The outcome measure – change in blood pressure – can be reliably measured and is subject to limited observer bias especially if there is adherence to methods for performing blood pressure measurements detailed in a study operations manual.

## ***5.5 Randomization and Concealment of Allocation***

Random, concealed allocation to intervention groups is, by definition, an essential component of an RCT and helps in minimizing any selection bias in the intervention groups under study.

Randomization refers to assignment of study subjects by chance to one of the groups in the clinical trial. This is done by using either a random number table or now more commonly a computer-generated random number list to assign sequential

study subjects to an intervention. The random number list links to treatment intervention in a consistent way. For instance, all even numbers on the random number list can be assigned to treatment A and odd numbers to treatment B. If there are three interventions, an option would be for treatment or intervention A to be assigned to all sequential numbers 1–33 on the random number list; intervention B to numbers 34–66, and intervention C to numbers 67–99 [20].

Interventions are allocated sequentially to patients enrolled in the study. Persons not involved in study enrollment or implementation must link random numbers and treatment groups. The sequence of study group assignments is concealed from persons enrolling patients in or carrying out the study. Absent such concealment from persons responsible for study enrollment, there is considerable potential for biasing the study results. Take the example of a study of an anti-hypertensive agent. Suppose the persons responsible for enrolling study subjects knew that the next allocated intervention was the drug they hoped to prove efficacious. In that case, they might discourage an otherwise eligible patient with potentially confounding conditions lessening therapeutic efficacy (obesity and history of smoking) from enrolling in the study. They would then wait to enroll a patient they thought was more likely to respond to their preferred intervention.

Concealment is done even when the study intervention cannot be masked by those conducting the study or providing care for the study subject. For instance, if the comparison was between drug therapy alone and exercise and weight loss programs to reduce hypertension, it is impossible to mask the intervention from those implementing the study and caring for study subjects. Even so, treatment allocation can and should be concealed.

## 5.6 *Methods of Randomization*

There are a number of ways in which randomization can be done. The simplest is *sequential randomization*. In this approach, there is a single sequential, random list for all subjects enrolled in the study. This works fine for large studies, where any chance of imbalances is likely minimal. In smaller studies, the chances of imbalance in assignment are greater (think of flipping a coin 20 times or 2000 times).

Using a *block randomization* method can minimize the risk of imbalance in study assignment. In a block randomization method, a block of defined numbers is identified in which there is an equal balance of study assignments. For instance, if a block size of six is chosen, and there are two study arms, three study subjects would be assigned to each treatment arm. This assures that even in small studies, there cannot be marked differences in subjects assigned to different treatment arms. A block randomization method also minimizes any temporal effects that might affect assessment of interventions.

A problem with block randomization is the potential to predict allocation, especially if it is not possible to blind interventions. Assuming a set block size of six, it can quickly become apparent to study personnel that there will be equal assignment

to the two study arms within each sequential six-study subject assignment. It would then become apparent that the next allocation would be to the latter group after a sequence in which three subjects have been assigned to drug therapy and two to weight loss and exercise.

A *permuted block randomization* method helps avoid this problem. In this method, the size of blocks varies. This diminishes but does not eliminate, study personnel from anticipating the next study allocation in an unblinded study with block randomization. As the size of blocks is likely to be limited, a good “card counter” could of course figure out the sequence of blocks. In practice, this is unlikely to happen.

There are additional permutations of block randomization. For multicenter studies, there can be separate randomization for each site. When there are especially important confounders that investigators want to control for, *stratified randomization* is used.

### **5.6.1 Community or Cluster Randomization**

Randomization by groups or clusters is used when the intervention is at group rather than individual level. To refer back to a blood pressure example. Suppose investigators aim to show the benefit of reducing blood pressure through a clinic-based public education effort, with outreach by community health workers. In that case, the level of randomization is likely to be the clinic, or a health district, rather than individuals. If the randomization was by community health workers rather than a clinic, there would likely be “contamination” because of the adjacency between persons. Some persons might not have a community health worker assigned to implement the public education effort, but their neighbors might have. Additionally, another community health worker at the same clinic assigned to implement the public health outreach effort might influence the community health worker not trained in the outreach program. To prevent such contamination, clinics, or even health districts, might be the unit of randomization. There are specific challenges when randomizing clusters [21] especially because the possibility of confounding and bias is likely to be much greater when the unit of randomization is a cluster. It is important to have knowledge of the communities or units to be included in the clusters and ensure that they are similar in terms of characteristics that might potentially confound the outcome. It is also important that the clusters are geographically distant enough to avoid a spillover effect.

### **5.6.2 Defining the Population Enrolled in the Study**

If a study’s results are widely applicable to others with the same condition, it is important to define the population enrolled in the trial. This includes the general population from whom study subjects were selected (hospital-based or clinic population, all consecutive eligible patients, or a convenience sample when study staff is

available to enroll patients). Inclusion and exclusion criteria need to be clearly defined. To continue the example of the study of an antihypertensive agent, the investigators need to specify the severity of the hypertension of enrolled patients; were morbidly obese persons or smokers excluded; the duration of preexisting hypertension; previous drug therapy; the gender of participants; their ethnicity; and their socioeconomic status. Clearly defining the population studied allows others to understand the potential utility and generalizability of the study findings.

## 6 Trial Organization and Management

Clinical trials require a clearly defined organizational structure, with each element of the organization having a defined set of responsibilities. The days when a fearless medical investigator had an inspired insight and went and tested that insight on patients without reference to any bureaucratic structures belongs to the era of black-and-white films and nineteenth century novels.

### 6.1 Study Sponsor

All clinical trials should have a defined sponsor. The sponsor is the “individual, company, institution, or organization that takes responsibility for initiating, managing, and/or financing a clinical study or trial” [22]. For commercially initiated and funded clinical trials, the company funding the trial is the sponsor. When noncommercial organizations – such as the United States National Institutes of Health – initiate and fund a clinical trial, they may serve as the sponsor. This is especially so for multicenter trials. For investigator-initiated noncommercial research, the investigator’s employer – most commonly a university or hospital – is usually the sponsor. The funding source is not always the study sponsor. Such as when a commercial entity funds university-based research initiated by university research staff. The sponsor has ultimate responsibility (and liability) for a clinical trial, ensuring that a trial is of sound quality – both scientifically and ethically.

Depending on the size and complexity of the study, clinical trials may be three levels of oversight and responsibility.

### 6.2 Trial Management Group

The first level of responsibility for conducting the study always lies with the investigative team members, who have day-to-day responsibility for conducting the study according to protocol and expeditiously. For large and complex studies, and especially for multicenter studies, a committee of investigators – variously termed the

“Trial Management Group” – is commonly established [23]. Such a committee consists of persons actively designing and conducting the study. There should be clearly defined standard operating procedures (SOPs) for the study, so that actions by all staff carrying out the study are consistent, and observations are recorded in a consistent manner.

### ***6.3 Data Monitoring and Safety Boards (DMSB, or Data Monitoring Committee)***

These are composed of persons independent of the study – i.e., not employed by the sponsor of the organization under whose aegis the study is being conducted and have no potential financial conflicts of interest posed by the interventions under study. The DMSB is responsible for reviewing interim or cumulative data for study-related adverse events; for evidence of efficacy before full enrollment in the study is completed; for quality of study data; timeliness of study enrollment and predicated completion; and protocol violations. Interim analysis to examine the efficacy of the intervention before planned study completion (or inferiority of the intervention) is done at pre-agreed intervals incorporated in the study protocol. The statistical analysis must account for the effect that multiple looks at study results will have the significance level required to assure that the results are unlikely to reflect chance alone (i.e., the more frequent examination of the results, the more likely that the null hypothesis will be rejected by chance alone).

### ***6.4 Trial Steering Committee***

The sponsor may delegate their senior-level oversight responsibility to a Trial Steering Committee. This committee includes both persons involved with study implementation and independent members, with one of the latter serving as the chair of the committee. Laypersons or representatives of the sponsor may serve on the committee. The committee regularly reviews the study’s progress to its objectives; receives reports and recommendations from the Data Monitoring and Steering Committee; and reviews information from external sources that may affect the study (such as other contemporaneous studies that definitively show toxicity or efficacy of the intervention under study). The Trial Steering Committee in consultation with the investigators and the sponsor makes decisions on the premature termination of the study (or prolongation beyond the expected completion date).

Not all studies, especially less well-funded smaller studies, will have this complexity of formal organizational structure and oversight. They should nonetheless adhere to all of the same principles for good clinical practice and assurance of ethical conduct. This is often done by preexisting structures in institutions, such as

ethical review committees (Institutional Review Boards) and preexisting scientific review committees.

Contract-research organizations, which commonly carry out studies on behalf of commercial sponsors, often establish a DMSB and Trial Management Committee or their equivalent [24]. When this is the case, conflicts may arise in larger studies involving other institutions, especially academic institutions, where the relationships and authority have not been clearly defined [25].

## 7 Data Analysis and Reporting

Data analysis and reporting of clinical trials should derive from a clearly defined hypothesis and trial objectives stated in the study proposal and protocol. A detailed analysis plan should be included in the proposal and adhered to upon study completion. The analysis should be conducted blindly – before it is known which intervention was provided to the groups under study.

In practice, this means that at the end of the study, a person not associated with the team implementing the study, or involved in the analysis, should identify the intervention assigned to each study subject. The persons conducting the analysis then do so without knowledge of the intervention received by each group.

Outcome measures should be clearly defined in the study protocol. As much as possible, the outcome measures should be objective, quantifiable, and not subject to interpretation by the study investigators. When an outcome measure is not easily quantifiable – or subject to a degree of interpretation – someone other than the study investigator should interpret any results pertinent to the study outcome.

There are several options as to which study subjects should be included in the analysis results. The commonly recommended *intention-to-treat analysis* includes all study subjects assigned to treatment intervention. No matter if they withdrew from the study before receiving any treatment, were noncompliant with the intended intervention, or left the study before the assessment of study endpoints.

The intention-to-treat analysis is conservative and likely to underestimate the effect of the intervention. This may especially be so when there are a large number of study subjects who never received the intervention or dropped out of the study before a final endpoint is reached. Especially for studies with a binary outcome where patients who left before the outcome is assessed are considered “treatment failures” for analysis.

A *per-protocol analysis* includes only those subjects who adhered fully to the proposed intervention. This most clearly resembles an effectiveness trial, which measures the beneficial effects of the intervention in real-world settings. In the context of a clinical trial, especially one in which the intervention cannot be blinded, per-protocol analysis lends itself to study bias by excluding investigators of study subjects they think are less likely to respond to the intervention. In a meta-analysis, per-protocol analyses showed a modestly greater treatment effect than intention-to-treat analyses [26].

A *modified intention-to-treat* analysis can be used, for instance, excluding study subjects who were never exposed to the intervention [27]. Sensitivity analyses can be conducted to assess the effect of missing data, including uncertain study outcomes, minor-protocol violations, or treatment effects within subgroup of patients [28]. Sensitivity analyses give an idea of the robustness of study findings, but should be clearly identified as a post hoc analysis. Another method is the *complier average causal effect (CACE)* analytic method, which has especially been used in adaptive intervention designs, where the study subject assigned to the intervention arm can decide what if any, intervention they accept [29]. CACE attempts to identify individuals in the control group who would have complied with the treatment given the opportunity to do so and uses this subset to compare to those in the intervention group accepting the intervention.

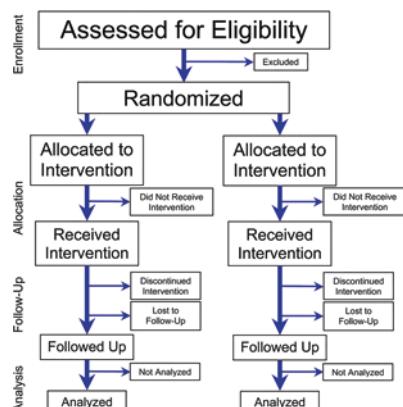
A standardized flowchart, such as the one developed by CONSORT (“Consolidated Standards of Reporting Trials”) shown below, should be used to report subject study participation (Fig. 6.1).

Because of the equipoise assumption underpinning most clinical trials – that there is true uncertainty about the superiority or inferiority of any of the study interventions – the most common underlying basis for analysis is the null hypothesis. The null hypothesis assumes that there is no true difference between the interventions. Statistical inference is used to determine if the null hypothesis is true or if it is rejected (i.e., one intervention performs differently than other interventions).

A basic assumption underlying the testing of the null hypothesis is the groups under study are drawn from the same population and are thus comparable. Hence, the need for random assignment to minimize variances between the two populations.

No matter how large the sample, two samples of the same population will rarely be the same. Chance alone will result in differences in the samples – even if the populations are very similar. Thus, statistical testing is needed to determine if the differences between the sampled populations under study – those receiving different interventions – occurred by chance alone or represent a true difference.

**Fig. 6.1** CONSORT flow diagram of the progress through the phases of a parallel randomized trial of two groups [30]



Different tests of statistical significance are used depending on the outcome measures being compared – continuous variables, for instance (the absolute value of systolic blood pressure) or a categorical outcome (diastolic blood pressure  $\geq 130$  mmHg). The statistical methods used to test these differences are described in detail in Chaps. 17, 21, and 22 of this book.

Several conventions govern the statistical inferences that are reported. One is that when differences are seen between groups, to reject the null hypothesis, we require that the differences between populations had only a 5% chance (a “*p*-value” of 0.05) of occurring by chance alone. Inversely, there is a 95% probability that the differences seen between groups represent a true difference. This type of error – invalidly rejecting the null hypothesis – is termed a Type 1 error. There is, however, nothing sacrosanct about a *P* value of 0.05. Using a more restrictive threshold for statistical inference is just as valid – for instance, a *P* value of 0.01. If a statistical significance test met this threshold, there would only be a one in one-hundred chance that the differences observed between groups occurred by chance alone. Conversely, a *P* value  $>0.05$  or 0.01 does not mean that is no possibility of benefit. It may be that larger studies are required or that subsequent studies may show an effect in similar or different populations.

A common way of expressing results in clinical trials is to show the 95% confidence intervals (CI) for the difference between the groups under study. If the 95% CIs (or 99% if one wants to be more restrictive) for the difference between groups do not include zero, then the null hypothesis is rejected, and the interventions being compared are assumed to be truly different in their effect on the population under study. 95% or 99% CIs are congruent with *P* values of 0.05 and 0.01.

Most major clinical journals have adopted standardized reporting of clinical trial results using the CONSORT guidelines [31]. These guidelines required reporting all relevant elements of a clinical trial, including the study title and abstract; introduction; methods; results; discussion; and other important information, including the source of funding and access to the study protocol. The guidelines specify the details required in each reporting section. This reporting format provides, along with the CONSORT study subject flow diagram, for consistent interpretation of study results. In addition to reporting parallel group randomized trials, CONSORT has guidelines for reporting cluster randomized trials, noninferiority and equivalence trials, nonpharmacological treatments, herbal interventions, and pragmatic trials. Using these standardized guidelines also enhances the feasibility of meta-analyses incorporating multiple studies.

## 8 Ethical Considerations and Informed Consent

There are common ethical principles underlying all human subject research, including clinical trials. These principles are detailed in a number of statements that underpin the ethical conduct of research. The first systematic and widely accepted set of principles was the Nuremberg Code, which was formulated in 1947 during the

trial of Nazi doctors who tortured and murdered prisoners and concentration camp inmates under the guise of doing medical research [32]. The Nuremberg Code emphasized the need for the voluntary consent of anyone who entered into a research study; the right of the study subject to withdraw from the study at any time; the need to minimize risk; to balance risk with the potential benefit of the experiment; and for a valid scientific premise for the research. The Nazi doctors who professed to conduct research did none of these things.

The Helsinki Declaration of the World Medical Association has further adapted and expanded upon the Nuremberg Code [33] as the ethical guidelines developed by the Council for International Organizations of Medical Sciences (CIOMS) in conjunction with the WHO [34]. The US government has also issued a set of principles and guidelines for the ethical conduct of research – the Belmont Report [35] – which highlighted three core ethical concepts – respect for persons, beneficence, and justice. This in turn guides ethical activities concerning informed consent, assessment of risks and benefits, and selection of research subjects.

Emanuel, Wendler, and Grady 2020 synthesized these guidelines and identified seven basic principles underpinning the ethical conduct of research [36]:

- (i) Value – the research must enhance health or knowledge
- (ii) Scientific validity
- (iii) Fair subject selection
- (iv) Favorable risk–benefit ratio
- (v) Independent review
- (vi) Informed consent
- (vii) Respect for enrolled subjects

Procedures and rules governing ethical research derive from these ethical precepts. In the United States, federal regulations (Code of Federal Regulations Title 45 Part 46) codify the rules governing research with human subjects [37]. These regulations have been adopted by most United States government departments and apply to all research conducted by United States government employees, research funded by the United States government, and institutions supported by the US government. Most other countries have similar regulations for the ethical conduct of research.

One of the fundamental elements arising from the Belmont report, and the Code of Federal Regulations, is the need for independent panels (termed Institutional Review Boards in the United States, and Ethical Review Committees in many locales) to review the ethics of all proposed clinical research. This ethical review has to be conducted along with creditable research review that assures the research's value and validity. The format and structure of the ethical review committees is established in the Code of Federal Regulations. Most institutions where research is conducted will have such review boards or will have a cooperative agreement with ethical review boards at other institutions.

**Table 6.2** Ethical responsibilities of a principal investigator in a clinical trial

Responsibility of Principal Investigator (PI)	Comment
Obtains approval for the study from the Institutional Review Board (IRB)	Most IRBs have a set format that is used for this application. Approval must be obtained before any subjects are enrolled in the study. The PI must be assured that the IRB meets the requirements for IRB as specified in 21 Code of Federal Regulation Part 56 [37]
Must conduct study according to the approved protocol	Changes to protocol need to be approved by IRB with notification to the sponsor
Takes responsibility for personally conducting or supervising the study	Virtually all studies – and especially large studies – involve a team of investigators and staff. The principal investigator and the sponsor are responsible for conducting the study according to all ethical guidelines
Assure all involved with the study, including all staff, are aware of ethical requirements.	This requires active educational efforts by the principal investigator, the study sponsor, and the IRB to ensure that staff has demonstrated knowledge of ethical requirements and their implementation
Assures that informed consent – written or oral – is properly obtained from study subjects	See details of informed consent requirements in Table 6.3
Report adverse events to the sponsor (and ensure that they are reported to IRB)	It is important to report anything that can be considered an adverse event. It must also ensure that the data monitoring and safety committee (if one was constituted) has information on adverse events available to it during its regular reviews
Maintain all study records and ensure they are accessible	This includes all patient documentation and original copies of the informed consent form if written informed consent is obtained on paper. If electronic forms are used, they must ensure that they are secure, remain confidential, and have robust systems to ensure that they remain accessible
Most report any potential conflict of interest to the sponsor and IRB	This particularly applies to financial conflicts and to all staff involved in the study

In the United States, principal investigators for drug studies aimed at obtaining an investigational new drug application (IND) must sign a legally binding Federal Drug Administration form – FDA 1572 – that specifies the responsibility of the investigator for the ethical conduct of the study [38]. Mandated responsibilities of the principal investigator are summarized in Table 6.2:

Consent is meaningless if it is simply a formality. Similarly, study subjects must be kept apprised of their progress and progress of the study. The major components of an informed consent form are as follows (Table 6.3).

**Table 6.3** Elements of informed consent form [37, 39]

Element required for informed consent	Comment
Language	The written form should avoid scientific or medical jargon and be written in language that can be comprehended by persons with basic literacy and education
Description of study	This should include a statement that the study involves research, the purposes of the research, the proposed duration of the study and duration of the subject's participation, and identification of interventions that are experimental. This should include the proposed number of persons to be enrolled in the study. If the study protocol is changed during the study in ways that may affect the study participant, additional consent should be obtained
Risks and discomfort	The focus should be on likely risks and discomforts, especially those that are potentially serious. An exhaustive detailed description of remote and unknown risks is likely to decrease the clarity of the consent form and lessen the understanding by the potential study participant
Benefits of participation in the study	This should include both potential benefits to the study subject and benefits that might accrue to others as a result of the study
Alternative interventions or treatments	If other interventions are available, this must be explained to the study subject. This should include the pros and cons of each option
Confidentiality	The study subject should be informed of who else besides study investigators and staff might have access to their records. It should also specify if their identifying information were available to those having access
Compensation and medical treatment in event of injury	The consent form should specify if compensation or treatment in the event of an adverse outcome or injury resulting from the study is available and how to access such support. Compensation or free medical care for study subjects in the event of an adverse outcome or injury resulting from clinical trials is not mandatory in the United States [40]. Compensation requirements also vary in other countries [41]
Voluntary participation	The form has to make clear that study participation is voluntary and that the study participant can withdraw from the study at any time without jeopardizing their right to the current standard of care
Contact information	The informed consent form must contain information on who the study subject can contact with questions and how to contact them. This contact information should be valid even after the study subject completes their direct participation in the study

**Box 2**

The study personnel involved in obtaining consent should not consider the informed consent process a formality or obstacle to overcome. Though there is always pressure to meet study enrollment targets, study staff have an ethical obligation to ensure that consent is truly informed and voluntary. As with all aspects of the ethical conduct of clinical trials, the onus falls upon the principal investigator and study staff to internalize the values of ethical behavior and act accordingly. No set of rules and guidelines alone can ensure the ethical conduct of studies absent staff embodying those values [37].

Distributive justice is an additional critical element of the ethical conduct of studies but one that is less amenable to formal guidelines and regulations. Rather than primarily involving conduct involving individual study subjects, it often deals with larger social issues – the “fair allocation of society’s benefits and burdens” [42]. Most of the statements on ethical principles of clinical research were in part motivated by clear violations of distributive justice. Disadvantaged members of society – the poor, prisoners, persons of color, and persons living in poor countries – bore the burden of potentially (or self-evidently) dangerous research without having access to the potential benefits of research.

The issue of distributive justice also extends to women and children, as clinical research and trials often excluded them. A larger issue of distributive justice is the allocation of research funds. Pharmaceutical companies drive much clinical research. Given the drive for profits, much of the research is aimed at lucrative segments of the pharmaceutical market – often drugs for chronic illnesses of older persons in rich countries. This has led to a proliferation of “me-too” studies aiming to identify drugs similar to already proven therapies [43]. Although there may be a benefit to some of these drugs, the larger question remains if some of these societal resources could more justly be directed to other conditions, especially those in resource-constrained countries where the potential societal benefit is much more significant.

## 9 Further Practice

1. Which of the following are **not** required elements of a clinical trial (choose all that apply):
  - (a) Involves human subjects
  - (b) Prospective enrollment
  - (c) Blinding of participants
  - (d) At least one comparison group
  - (e) Randomization of study participants to different interventions
2. Match the phase of the clinical study to what is done in the phase of the study:  
Study phase: 1; Study phase: 2; Study phase: 3; Study phase: 4.  
Match phases 1 through 4 with:
  - (a) Preliminary testing of efficacy and dosing amount and more detailed safety information
  - (b) A long-term study of side effects and benefits after licensing a drug, biological product, or device
  - (c) To evaluate the safety of a drug, biological product, or device
  - (d) Comparative study of the efficacy of intervention; compare with standard treatment (or placebo)

3. Equipoise in clinical trials refers to which of the following:
  - (a) The correct matching of intervention and control groups
  - (b) The lack of bias in the selection of participants
  - (c) The lack of bias in data analysis
  - (d) Lack of evidence of the superiority of any of the trial interventions
4. Double-blinding in a clinical trial refers to the following:
  - (a) The study participant does not know the identity of the drug used
  - (b) The investigators do not know the identity of the drug used
  - (c) Both the participant and the investigators do not know the identity of the drug used
5. Blinding of a clinical trial reduces
  - (a) Bias
  - (b) Confounding
  - (c) Both
6. A problem with block randomization is the potential to predict treatment allocation – True or False
7. The study sponsor is the entity that is responsible for:
  - (a) Carrying out the day-to-day conduct of a clinical trial
  - (b) Analyzing the study results
  - (c) Initiation, management, and/or financing of a clinical study or trial
  - (d) Writing the report of the study
8. An intention-to-treat analysis excludes which of the following groups:
  - (a) Study subjects who do not complete the study
  - (b) Study subjects who never received the study intervention
  - (c) Study subjects that completed the study but did not adhere to the study protocol
  - (d) None of the above
9. The null hypothesis assumes:
  - (a) That groups being compared can never be the same
  - (b) The groups being compared will only differ because of sampling error
  - (c) That there is no statistically significant difference between the groups being compared
10. Ethical responsibilities of a clinical trial principal investigator include (select all that apply):
  - (a) Obtain approval for the trial from an authorized, ethical review committee
  - (b) Conduct the study according to the approved protocol
  - (c) Decide if a study should be stopped because of the frequency of adverse reactions to one of the study interventions

- (d) Have a responsibility to reimburse study subjects if they suffer harm from the study intervention
- (e) Assures that informed consent is obtained from study subjects

### Answer Keys

1. (c) and (e)
2. 2 phase 1 matches to c; phase 2 matches to b; phase 3 matches to d; phase 4 matches to b
3. (a)
4. (c)
5. (a)
6. True
7. (c)
8. (d)
9. (c)
10. (a), (b) and (e)

## References

1. Shaw GB. The doctors dilemma: preface on doctors. Salt Lake City: Project Gutenberg Literary Archive Foundation; 1909.
2. Dixon JR Jr. The international conference on harmonization good clinical practice guideline. Qual Assur. 1998;6(2):65–74.
3. Committee for Human Medicinal Products. Guideline for good clinical practice E6(R2). In: European Medicines Agency, editor. London, UK: European Union; 2016.
4. World Health Organization. Handbook for good clinical research practice (GCP): guidance for implementation. Geneva: World Health Organization; 2005.
5. Greene JA, Podolsky SH. Reform, regulation, and pharmaceuticals--the Kefauver-Harris amendments at 50. N Engl J Med. 2012;367(16):1481–3.
6. NIH. NIH's Definition of a Clinical Trial 2023. 7 January 2023. Available from: <https://grants.nih.gov/policy/clinical-trials/definition.htm>.
7. Lobanovska M, Pilla G. Penicillin's Discovery and antibiotic resistance: lessons for the future? Yale J Biol Med. 2017;90(1):135–145.
8. Vecchio I, Tornali C, Bragazzi NL, Martini M. The discovery of insulin: An important milestone in the history of medicine. Front Endocrinol (Lausanne). 2018;9:613. <https://doi.org/10.3389/fendo.2018.00613>.
9. Bédécarrats F, Guérin I, Roubaud F, editors. Randomized control trials in the field of development a critical perspective. Oxford, UK: Oxford University Press; 2020.
10. Beaton A, Okello E, Rwebembeira J, Grobler A, Engelman D, Alepere J, et al. Secondary antibiotic prophylaxis for latent rheumatic heart disease. N Engl J Med. 2022;386(3):230–40.
11. SPIRIT. SPIRIT (Standard Protocol Items: Recommendations for Interventional Trials). Available from: <https://www.spirit-statement.org/>.
12. U.S. National Library of Medicine. ClinicalTrials.gov 2023. Available from: <https://clinicaltrials.gov/>.
13. EudraCT. (European Union Drug Regulating Authorities Clinical Trials Database 2023 [cited 2023 January 7]. Available from: <https://eudract.ema.europa.eu/>.
14. EDCTP. Pan-African Clinical Trials Registry 2023. Available from: <http://www.edctp.org/pan-african-clinical-trials-registry/>.

15. IC RTP Registry Network. International Clinical Trials Registry Platform (ICTRP). Available from: <https://www.who.int/clinical-trials-registry-platform/network>.
16. London AJ. Equipoise and international human-subjects research. *Bioethics*. 2001;15(4):312–32.
17. Karlawish JH, Lantos J. Community equipoise and the architecture of clinical research. *Camb Q Healthc Ethics*. 1997;6(4):385–96.
18. Hrobjartsson A, Thomsen AS, Emanuelsson F, Tendal B, Hilden J, Boutron I, et al. Observer bias in randomised clinical trials with binary outcomes: systematic review of trials with both blinded and non-blinded outcome assessors. *BMJ*. 2012;344:e1119.
19. Moustgaard H, Clayton GL, Jones HE, Boutron I, Jorgensen L, Laursen DRT, et al. Impact of blinding on estimated treatment effects in randomised clinical trials: meta-epidemiological study. *BMJ*. 2020;368:l6802.
20. Altman DG, Bland JM. How to randomise. *BMJ*. 1999;319(7211):703–4.
21. Esserman D, Allore HG, Travison TG. The method of randomization for cluster-randomized trials: challenges of including patients with multiple chronic conditions. *Int J Stat Med Res*. 2016;5(1):2–7.
22. United Kingdom National Health Service Health Research Authority. UK policy framework for health and social care research. London, UK: UK National Health Service; 2017.
23. Harman NL, Conroy EJ, Lewis SC, Murray G, Norrie J, Sydes MR, et al. Exploring the role and function of trial steering committees: results of an expert panel meeting. *Trials*. 2015;16:597.
24. Beach JE. Clinical trials integrity: a CRO perspective. *Account Res*. 2001;8(3):245–60.
25. Gibson CM, Goldhaber SZ, Cohen AT, Nafee T, Hernandez AF, Hull R, et al. When academic research organizations and clinical research organizations disagree: processes to minimize discrepancies prior to unblinding of randomized trials. *Am Heart J*. 2017;189:1–8.
26. Mostazir M, Taylor G, Henley WE, Watkins ER, Taylor RS. Per-protocol analyses produced larger treatment effect sizes than intention to treat: a meta-epidemiological study. *J Clin Epidemiol*. 2021;138:12–21.
27. White IR, Carpenter J, Horton NJ. Including all individuals is not enough: lessons for intention-to-treat analysis. *Clin Trials*. 2012;9(4):396–407.
28. Thabane L, Mbuagbaw L, Zhang S, Samaan Z, Marcucci M, Ye C, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Med Res Methodol*. 2013;13:92.
29. Connell AM. Employing complier average causal effect analytic methods to examine effects of randomized encouragement trials. *Am J Drug Alcohol Abuse*. 2009;35(4):253–9.
30. Consort - Consolidated Standards of Reporting Trials. Consort flow diagram 2010. Available from: <https://www.consort-statement.org/consort-statement/flow-diagram>.
31. Schulz KF, Altman DG, Moher D, Group C. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *BMJ*. 2010;340:c332.
32. Shuster E. Fifty years later: the significance of the Nuremberg code. *N Engl J Med*. 1997;337(20):1436–40.
33. World Medical Association. WMA declaration of Helsinki – ethical principles for medical research involving human subjects. Available from: <https://www.wma.net/policies-post/wma-declaration-of-helsinki-ethical-principles-for-medical-research-involving-human-subjects/>.
34. Council for International Organizations of Medical Sciences (CIOMS). International ethical guidelines for biomedical research involving human subjects. Geneva, Switzerland; 2016.
35. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. Ethical principles and guidelines for the protection of human subjects of research. Washington, DC: United States Department of Health, Education, and Welfare; 1979. Available from: <https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html>
36. Emanuel EJ, Wendler D, Grady C. What makes clinical research ethical? *JAMA*. 2000;283(20):2701–11.

37. United States Department of Health and Human Services. Policy for protection of human subjects. Washington, DC; 2023. Available from: <https://www.ecfr.gov/current/title-45 subtitle-A/ subchapter-A/part-46>.
38. United States Food and Drug Administration. Frequently asked questions – statement of investigator (Form FDA 1572). Guidance for Sponsors, Clinical Investigators, and IRBs. 2010. Available from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/ frequently-asked-questions-statement-investigator-form-fda-1572>.
39. United States Food and Drug Administraiton. Informed consent: draft guidance for IRBs, clinical investigators, and sponsors Washington, DC. 2014. Available from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/informed-consent#genrequirments>.
40. Resnik DB, Parasidis E, Carroll K, Evans JM, Pike ER, Kissling GE. Research-related injury compensation policies of U.S. research institutions. IRB. 2014;36(1):12–9.
41. Chingarande GR, Moodley K. Disparate compensation policies for research related injury in an era of multinational trials: a case study of Brazil, Russia, India, China and South Africa. BMC Med Ethics. 2018;19(1):8.
42. The National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. Ethical principles and guidelines for the protection of human subjects of research: United States Department of Health and Human Services; 1979. Available from: <https://www.hhs.gov/ohrp/regulations-and-policy/belmont-report/read-the-belmont-report/index.html>.
43. Aronson JK, Green AR. Me-too pharmaceutical products: history, definitions, examples, and relevance to drug shortages and essential medicines lists. Br J Clin Pharmacol. 2020;86(11):2114–22.

# Chapter 7

## Screening



Amal K. Mitra

### Learning Objectives

After completing this chapter, you will be able to:

- Evaluate characteristics of a screening test
- Describe commonly used screening tests including genetic screening
- Apply evaluation tools of screening tests using real-life examples
- Understand how people can be misclassified into screening test positive or negative
- Find the use of an ROC curve

## 1 Introduction

Screening is a tool that is used in both clinical practice and in public health. Clinicians use a screening tests for their patients for an early detection of a disease so that treatment can be initiated soon. However, screening test is not a diagnostic test. Screening tests screen out apparently healthy individuals to find out the likelihood of a suspected disease. For example, your doctor advised you to perform a colonoscopy. Colonoscopy is a screening test to rule out colon cancer. By doing a colonoscopy test, your doctor may not find any abnormalities in the entire colon. Then he may ask you to come for another test after 5 years. On the other hand, your doctor may find a few suspected precancerous nodules or cysts. He will do another test called biopsy from the sample to confirm his suspicion. Biopsy is a confirmatory diagnostic test. After a screening test, we need to do one or more confirmatory

---

A. K. Mitra (✉)

Department of Epidemiology and Biostatistics, Jackson State University, Jackson, MS, USA  
e-mail: [amal.k.mitra@jsums.edu](mailto:amal.k.mitra@jsums.edu)

tests for the diagnosis and treatment. The screening test is evaluated against a confirmatory diagnostic test to evaluate how good the screening test is in terms of the confirmatory test, called “gold standard.”

A public health specialist will advise you to do a screening test for the same purpose of early diagnosis of a disease. In public health, screening test falls under one of the three stages of disease prevention – it is called “secondary prevention.” By secondary prevention, a public health specialist aims to prevent a disease before it is severe enough to cause more complications or a worse prognosis. To further clarify, by the term “secondary prevention,” we mean the disease process or the pathology of the disease has already been in place, we can’t stop the disease from happening but we can probably slow down the disease sequelae by an early diagnosis and the initiation of early treatments. In this chapter, we will discuss all about screening tests.

## 2 Criteria for an Effective Screening Test

Several characteristics of diseases as well as the characteristics of the screening test make it effective.

### 2.1 Disease Characteristics

An ideal situation of conducting a screening test dependents on the following disease conditions:

#### 2.1.1 High Morbidity and Mortality

There are many diseases that are endemic in nature and usually not virulent enough that needs an early detection. For example, endemic flu in the United States. Flu is a seasonal disease, which lasts from October to May, with a peak activity from December to March. Although flu may cause pneumonia and other complications, most cases of flu never lead to pneumonia. Again, screening test is not needed when people already have clinical symptoms. On the other hand, there are many diseases that cause significant morbidity and mortality and are often detected when it is too late. In those cases, screening plays an important role. Example of diseases or conditions and the screening tests for the conditions are listed here:

- Breast cancer – Mammogram, magnetic resonance imaging (MRI), breast self-exam (BSE), biopsy
- Cervical cancer – Pap smear and the human papilloma virus (HPV) tests
- Prostate cancer – Prostate-specific antigen (PSA) test
- Colon cancer – Colonoscopy

- Diabetes – Blood sugar (fasting), hemoglobin A<sub>1</sub>C, glucose tolerance test
- Hypertension – Monitoring blood pressure
- High cholesterol – Monitoring lipid profile

### 2.1.2 Early Detection Helps in Prognosis

As mentioned earlier, some chronic diseases such as cancer are often detected at a late stage of the disease. Lung cancer is such a disease that makes challenge for early detection because the initial symptoms of the disease are nonspecific [1]. The majority of patients with lung cancer are diagnosed as a late stage with a poor prognosis of the disease. Researchers are looking for screening tests that can detect the disease early enough so that interventions can be initiated early for a better prognosis of the disease.

### 2.1.3 Availability of Treatment

Screening test is used for a disease for which available treatments can either cure the disease or minimize sufferings.

## 2.2 Ideal Characteristics of a Screening Test

The following characteristics should be considered in using a screening test.

- Valid and reliable
- High sensitivity
- High specificity
- Convenient (or noninvasive)
- Low cost
- Acceptable by the population

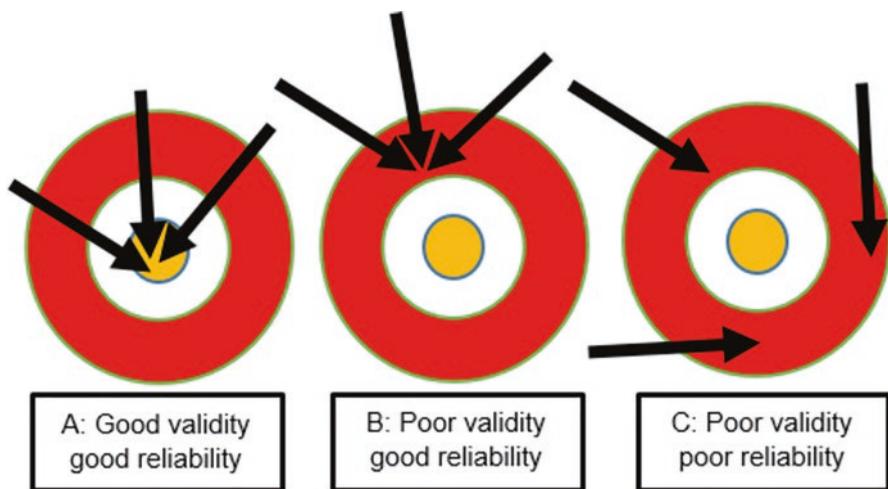
### 2.2.1 Validity and Reliability

Validity: It is the measurement of the truth.

Reliability: It is the reproducibility of the test. Reproducibility occurs when the measuring instrument repeatedly produce the same result, or when several observers get the same result after repeated tests using the same method or instrument.

Validity and reliability of a screening test can be illustrated by the scenario of a game called “dirt” (Fig. 7.1). In a dirt game, you throw arrows and your aim is to hit the bull’s eye to get the maximum points.

Figure 7.1a represents good validity and reliability because all the arrows hit the center, which means it finds the real true value. The figure in the middle (Fig. 7.1b)



**Fig. 7.1** Validity and reliability [2]. (Used with permission of John Wiley & Sons, Inc. from Chapter 15: Identifying Diseases by Screening, Mitra AK, first edition, 2023; permission conveyed through Copyright Clearance Center, Inc.)

shows that all the arrows hit at one point but they do not reach the center. This means the test is reliable but it is not valid. The figure to the right (Fig. 7.1c) represents neither validity nor reliability, as the arrows reach at different points, and they are away from the center.

### 2.3 Sensitivity

It is the proportion of the people who are screening test positive among the people who have the disease. Sensitivity relates to the test's ability to identify positive results. A test with high sensitivity can be considered as a reliable test. The formula can be illustrated by using a  $2 \times 2$  contingency table (Table 7.1).

$$\text{Sensitivity} = \frac{A}{(A + C)}$$

### 2.4 Specificity

Specificity is the proportion of the people who are screening test negative among the people who do not have the disease. Specificity relates to the test's ability to identify negative results.

**Table 7.1** Sensitivity, specificity, predictive values, and agreement of a screening test

Screening test	Population		Total
	With disease	Without disease	
Positive	Disease is present, and the test is positive = True positive (TP) ( <i>A</i> )	Disease is absent, but the test is positive = False positive (FP) ( <i>B</i> )	<i>A + B</i>
Negative	Disease is present, but the test is negative = False negative (FN) ( <i>C</i> )	Disease is absent, and the test is negative = True negative (TN) ( <i>D</i> )	<i>C + D</i>
Total	<i>A + C</i>	<i>B + D</i>	<i>A + B + C + D</i>

$$\text{Specificity} = \frac{D}{(B + D)}$$

## 2.5 Positive Predictive Value

It is the proportion of people who got the disease among those who are screening test positive.

$$\text{PPV} = \frac{A}{(A + B)}$$

## 2.6 Negative Predictive Value

It is the proportion of people who do not have the disease among those who are test negative.

$$\text{NPV} = \frac{D}{(C + D)}$$

## 2.7 Agreement of the Test

Total agreement of a screening test is the sum of true positive and true negative (*A + D*).

$$\text{Percent agreement} = \frac{(A + D)}{\text{Grand total}(A + B + C + D)}$$

### 3 Several Other Terms in Relation to Screening

Let us take a hypothetical model of cervical cancer to illustrate several terms related to a screening test (Fig. 7.2).

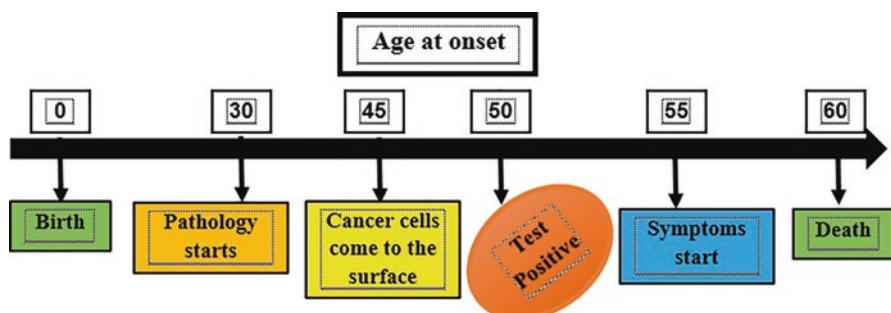
In case of cervical cancer, about 80–85% are squamous cell carcinoma, and most of the rest are adenocarcinoma. In the hypothetical model (Fig. 7.2), a woman suffered from cervical cancer. Suppose, in her case, the pathology of cervical cancer starts at age 30. It means cells in the cervix, the lower part of the uterus, start to become abnormal. Changes in the cell DNA make them to multiple out-of-control and they accumulate in growths called tumor cells. At age 45, cells from the surface of the cervix start to come out to the surface of the cervix – the process is called exfoliation. Only at this stage precancerous cells are detectable by cervical tissue biopsy.

In this hypothetical case, suppose the screening test is done at age 50 and clinical symptoms of cervical cancer begin at age 55. Let us define a few terms based on this age of onset of different stages of cervical cancer of the woman.

#### 3.1 Total Preclinical Phase (TPCP)

The time period from the beginning of the pathology until the appearance of clinical symptoms. In the above example, the pathology of neoplasm started at age 30 and the symptoms appeared at age 55.

$$\text{TPCP} = 55 - 30 = 25 \text{ years.}$$



**Fig. 7.2** A hypothetical model of cervical cancer

### 3.2 Detectable Preclinical Phase (DPCP)

The time period from the time the disease could be detected by the available screening test before the appearance of clinical symptoms. The reason that the screening test could be applied during this time is because the cells already started to come out or exfoliate to the surface of the cervix. In the above hypothetical case,

$$\text{DPCP} = 55 - 45 = 10 \text{ years.}$$

For a disease, if DPCP is long (as long as TPCP), then the screening test could detect the disease much earlier (when the cancer cells begin to come out to the surface of the cervix).

### 3.3 Lead Time

Time between the diagnosis by screening test and the appearance of clinical symptoms. In the above example (Fig. 7.2), the screening test became positive at age 50 and symptoms appeared at age 55.

$$\text{Lead time} = 55 - 50 = 5 \text{ years.}$$

### 3.4 Lead-Time Bias

Lead-time bias is a type of information bias specific to screening studies [3]. This bias occurs with a presumption that the patient survived longer because the screening was done earlier before the appearance of clinical symptoms.

Let us take another example of two women with breast cancer. Woman A went to a routine breast cancer screening program. She was diagnosed having breast cancer at age 55 by the screening test. Woman B went to a doctor because of sudden pain in her breasts, and she was diagnosed of breast cancer at age 60. Suppose, both died at 70.

Survival time for woman A, from the time of screening to death =  $70 - 55 = 15$  years.  
Survival time for woman B, from the time of doctor's visit to death =  $70 - 60 = 10$  years.

Looking at the above scenarios, lead time bias artificially increases the survival time of a patient who went through screening. However, both women may have started developing the disease at a similar time and both also died around the same age. It seems like breast cancer screening program increased the survival of woman A by  $(15 - 10) 5$  years compared to woman B.

To avoid lead time bias, you should not compare survival rates, rather you should analyze mortality rates of the population.

Example:

$$\text{Mortality rate of breast cancer} = \frac{\text{Number of deaths from breast cancer}}{\text{Total cases of breast cancer in a given year}} \times 1000.$$

### **3.5 Receiver Operating Characteristic (ROC) Curve**

#### **Box 1**

The area under the ROC curve (*AUC*) is a very widely used measure of performance for classification and diagnostic rules. To produce an ROC curve, the sensitivities and specificities for different values of a continuous test measure are first tabulated [4]. Then, the graphical ROC curve is produced by plotting sensitivity (true positive rate) on the y-axis against 1-specificity (false positive rate) on the x-axis for the various values tabulated. The area under the ROC curve (*AUC*) is a global measure of the ability of a test to discriminate whether a specific condition is present or not present. An *AUC* of 0.5 represents a test with no discriminating ability (i.e., no better than chance), while an *AUC* of 1.0 represents a test with perfect discrimination.

In Fig. 7.3, the point of the curve which is nearest to the top left hand corner represents the optimal compromise between sensitivity and specificity.

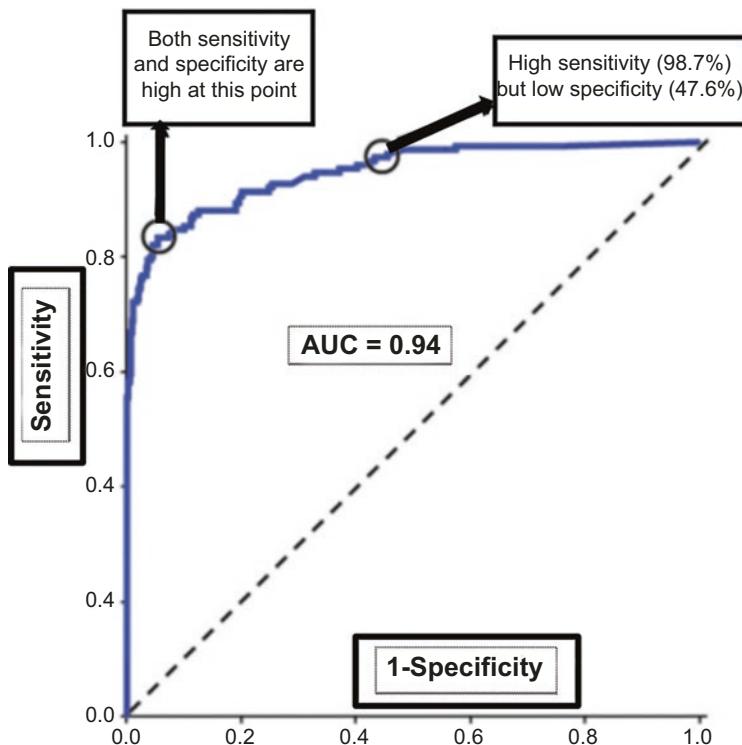
## **4 Commonly Used Screening Tests**

### **4.1 Mammogram**

This is the most commonly used screening test for breast cancer. The United States Preventive Services Task Force (USPSTF) recommends biennial mammogram for women aged 50–74 years and those who are at average risk for breast cancer. The decision of getting a mammogram for a woman before age 50 should be advised by the doctor.

### **4.2 MRI of Breast**

Breast MRI is used for women who are at high risk for getting breast cancer. Because of false-positive results, breast MRIs are not used for women who are at average risk.



**Fig. 7.3** Example of an ROC curve using hypothetical data

#### 4.3 Pap Smear

The test is named after a famous Greek doctor Georgios Papanikolaou. Women should start Pap smear screening of the uterine cervix at age 21. If the Pap smear test turns to be normal, it is repeated every 3 years. The Pap test can find abnormal cells in the cervix which may turn into cancer. The test may also detect infections of human papillomavirus (HPV), the most common infection causing cervical cancer and cancer of the adjacent parts of the uterus.

#### 4.4 Colonoscopy

A colonoscopy is a test to detect changes in the large gut such as polyps or growths that are suspected to be precancerous. Regular screening, beginning at age 45, is the key to preventing colorectal cancer and finding it early. The USPSTF recommends that adults age 45–75 be screened for colorectal cancer.

## **4.5 Low-Dose Computed Tomography (LDCT)**

A low-dose CT scan is a quick, painless, and noninvasive test to screen for lung cancer. According to the USPSTF, the following persons should do a yearly lung cancer screening with LDCT:

- Smoking history: (1) Have a 20 pack-year or more smoking; (2) smoke now or have quit within the past 15 years, and
- Age: Between 50 and 80 years old

## **4.6 PSA for Prostate Cancer**

Prostate-specific antigen (PSA) is a blood test for men to help detect prostatic cancer. People who are at a higher risk of prostate cancer, such as Blacks and family history (father or brother) of prostate cancer, may start PSA test as early as age 40–45. Otherwise, you may discuss with your doctor a PSA test if you are aged 55–69 years. The USPSTF does not recommend PSA test screening for men 70 years and older.

## **4.7 Genetic Screening**

Genetic screening is a study of a person's DNA to detect genetic differences or susceptibility to particular disease that may run in the family. It is also done to learn about the chance if a current or future pregnancy will have a genetic condition, or to diagnose a genetic condition if you or your child has symptoms. Most commonly, genetic screening includes tests such as phenylketonuria (PKU), Down syndrome, cystic fibrosis, congenital adrenal hyperplasia, and congenital hypothyroidism.

# **5 Evaluating Screening Tests: An Exercise**

Let us calculate some of the indicators for evaluating screening tests by using hypothetical data presented in Table 7.2. The screening test is evaluated against a specific confirmatory diagnostic test or gold standard (such as a biopsy for cancer or blood culture for a bacterial disease).

**Table 7.2** Data of screening test and the gold standard

Screening test	Population		Total
	With disease	Without disease	
Positive	50	100	150
Negative	30	820	850
Total	80	920	1000

True positive = both screening test and disease are positive = 50  
 False positive = screening test positive but disease negative = 100  
 False negative = screening test negative but disease positive = 30  
 True negative = both screening test and disease are negative = 820

$$\text{Sensitivity} = \frac{50}{80} \times 100 = 62.5\%$$

$$\text{Specificity} = \frac{820}{920} \times 100 = 89.1\%$$

$$\text{Positive Predictive Value} = \frac{50}{150} \times 100 = 33.3\%$$

$$\text{Negative Predictive Value} = \frac{820}{850} \times 100 = 96.5\%$$

Total agreement = True positive + True negative = 50 + 820 = 870

$$\text{Percent agreement of the test} = \frac{870}{1000} \times 100 = 87\%$$

## 6 Selecting a Cutoff Point

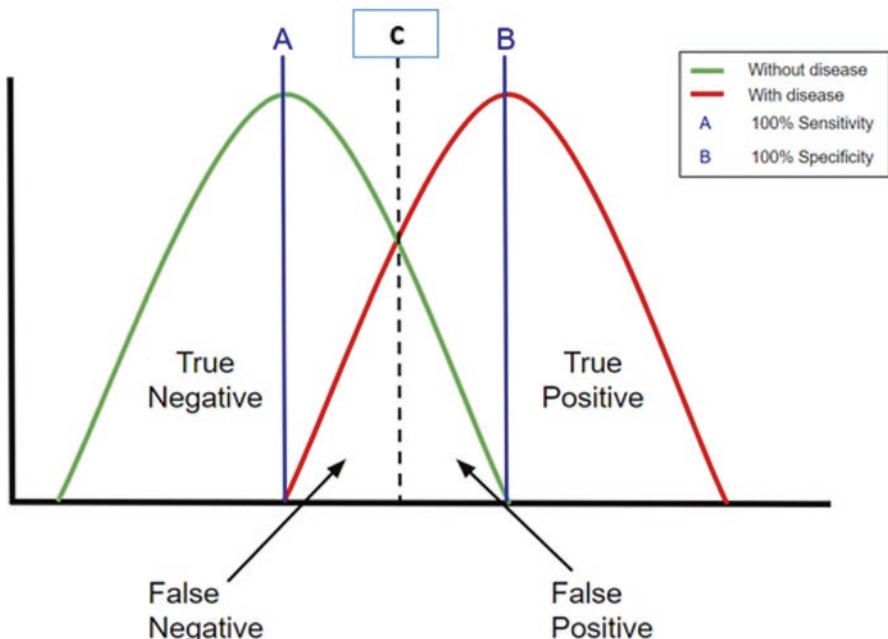
### Box 2

It is sometimes a dilemma how to use a cutoff point of a test. The decision is a tradeoff between sensitivity and specificity of the test. In other words, this depends on whether you want to improve sensitivity or specificity of the test. An increase in true positive will increase sensitivity, and an increase in true negative will increase in specificity. In some diseases, such as cancer, you want to be very specific in the diagnosis – i.e., very few people on a screening test should be false positive for cancer screening. The specificity will be high. In another situation, for example, COVID-19, you want to have a very sensitive screening test – i.e., the screening test should have very few false-negative cases.

Figure 7.4 demonstrates the effect of choosing various cutoff points between “clearly positive” and “clearly negative” test results in case of blood sugar test.

In Fig. 7.4, you maximize both sensitivity and specificity when you select the cutoff point at “C”. On the other hand, if you move the cutoff point to “A” by

## Sensitivity vs. Specificity



**Fig. 7.4** Effect of changing the cutoff point on sensitivity and specificity of the test [2]. (Used with permission of John Wiley & Sons, Inc. from Chapter 15: Identifying Diseases by Screening, Mitra AK, first edition, 2023; permission conveyed through Copyright Clearance Center, Inc.)

lowering down the blood glucose level that is to be classified as abnormal, almost all the individuals who have diabetes will have the screening test positive. In that case, sensitivity will be close to 100%, and specificity will be decreased because many of the nondiabetic individuals will be classified as having diabetes. Conversely, if you move the cutoff point of blood glucose at a higher level to “B,” specificity of the test will be increased to almost 100%. In the latter case, sensitivity will be compromised because many diabetic individuals will be misclassified as normal. Therefore, it is always a judgment call whether you want to increase the sensitivity or the specificity of the screening test.

## 7 Further Practice

### A. Short questions

1. You want all the individuals who are found screening test positive must have the disease. In this case, which one do you want to increase – sensitivity or specificity?

2. You want to have most of the individuals who are screening test negative also should not have the disease. Do you increase sensitivity or specificity of the test?
3. Screening test is a trade-off between sensitivity and specificity – please explain.
4. What is the recommendation for doing a PSA test above age 69?
5. Who should be screened for mammogram?
6. What is the screening test used for lung cancer detection?
7. A screening test is not a diagnostic test – please explain.
8. What age is not recommended for PSA test for prostate cancer?
9. What does a Pap smear test do for you?
10. Several polyps are found on a colonoscopy test. What is the next step in the diagnosis?

**B. Problem-solving questions**

Given the following data, please calculate:

1. Sensitivity
2. Specificity
3. Positive predictive value
4. Negative predictive value
5. Percent agreement of the test

Screening test	Population		Total
	With disease	Without disease	
Positive	40	60	100
Negative	90	110	200
Total	130	170	300

**C. Multiple choice questions**

1. Screening is a diagnostic test.      True/False
2. Sensitivity is high means the true positive is also high      True/False
3. Specificity is low means the false positive is high      True/False
4. Total agreement is sum of true positive and true negative      True/False
5. You can have a screening test with 100% sensitivity and 100% specificity      True/False
6. The following is true, EXCEPT,
  - (a) True positive means both test positive and disease positive
  - (b) True negative means both test negative and disease negative
  - (c) False positive means test negative and disease positive
  - (d) False negative means test negative and disease positive
7. The following answers are correct, EXCEPT,
  - (a) All the arrows of a dart game hit at one point and at the center – valid test
  - (b) All the arrows of a dart game hit at one point and at the center – reliable test

- (c) All the arrows of a dirst game hit at one point but not at the center – valid test
  - (d) All the arrows of a dirst game hit at one point but not at the center – reliable test
8. An ideal screening test should have the following (choose the correct answer):
- (a) Valid and reliable
  - (b) High sensitivity
  - (c) High specificity
  - (d) Convenient (or noninvasive)
  - (e) Low cost
  - (f) All of the above
9. High DPCP is a characteristic of (choose the correct answer)
- (a) Disease
  - (b) Screening test
  - (c) Both
  - (d) None
10. Lead time bias occurs when (choose the correct answer)
- (a) Screening test is done before the clinical diagnosis
  - (b) Screening test is not confirmed by a diagnostic test
  - (c) Screening test is done following the clinical diagnosis
  - (d) Screening is necessary but it is not done because the patient is biased

### **Answers Keys**

Section A: Review the chapter.

Section B: 1 – 30.8%; 2 – 64.7%; 3 – 40%; 4 – 55%; 5 – 50%.

Section C: 1 – False; 2 – True; 3 – True; 4 – True; 5 – False; 6 – c; 7 – c; 8 – f; 9 – a; 10 – a.

## **References**

1. Walter F, Rubin G, Bankhead C, et al. Symptoms and other factors associated with time to diagnosis and stage of lung cancer: a prospective cohort study. *Br J Cancer*. 2015;112(Suppl 1):S6–S13. <https://doi.org/10.1038/bjc.2015.30>.
2. Mitra AK. Screening. In: Epidemiology for dummies, epidemiology for dummies. Hoboken: Wiley; 2023.
3. Duffy SW, Nagtegaal ID, Wallis M, Cafferty FH, Houssami N, Warwick J, et al. Correcting for lead time and length bias in estimating the effect of screen detection on cancer survival. *Am J Epidemiol*. 2008;168(1):98–104. <https://doi.org/10.1093/aje/kwn120>.
4. Hoo ZH, Dandlish J, Teare D. What is an ROC curve? *Emerg Med J*. 2017;34:357–8. <https://doi.org/10.1136/emermed-2017-206735>.

# Chapter 8

## Surveillance: The Role of Observation in Epidemiological Studies



**Adetoun F. Asala**

### Learning Objectives

After completing this chapter, you will be able to:

- Understand public health surveillance and significance of surveillance
- Approaches to public health surveillance
- Differentiate between public health surveillance and epidemiologic research
- Describe the types and process of surveillance
- Describe public health surveillance systems

## 1 Introduction

Public health surveillance involves the continuous and ongoing systematic collection, analysis, and interpretation of relevant health data, which play key role in planning, implementing, and evaluating public health policies and practices as well as disseminating information needed for disease prevention and control [1, 2]. According to the Centers for Disease Control and Prevention (CDC), public health surveillance is the chief cornerstone of the entire public health practice; health planning, policy changes, and decision making is hinged on information made available through surveillance data [3]. Surveillance data are collected and collated through health surveys (National Health and Nutrition Examination Survey – NHANES), vital records (death, birth), registries (immunization and disease), electronic health records or information systems (utilization data and hospital discharge data), environmental monitoring, clinical and public health research, as well as other valuable data sources. The main goal of surveillance is to provide valuable information useful for disease prevention and control, thus enhancing data-driven decision making by monitoring disease burden over time, detecting changes in disease occurrence, and determining disease risk factors and populations at greatest risk [4]. Surveillance in public health dated as far back as the eighteenth century where John Snow who

---

A. F. Asala (✉)

Mississippi State Department of Health, Office of Preventive Health, Ridgeland, MS, USA  
e-mail: [Adetoun.asala1@msdh.ms.gov](mailto:Adetoun.asala1@msdh.ms.gov)

is regarded as the father of epidemiology solved a seemingly mysterious infectious disease then, cholera. He used basic surveillance principles to identify the cause and course of the disease. John Snow by surveilling the course and path of cholera in clusters within and around affected communities, he identified what was common around these areas, a particular water pump supplying water to the communities. Then he was able to end the spread of cholera disease without developing a treatment at that time [5].

In recent time, the importance of surveillance in epidemiological studies cannot be overemphasized, the COVID-19 pandemic even underscores the need to establish strong public health surveillance system especially in developing, middle- and low-income countries. Some highlighted benefits of surveillance in public health include:

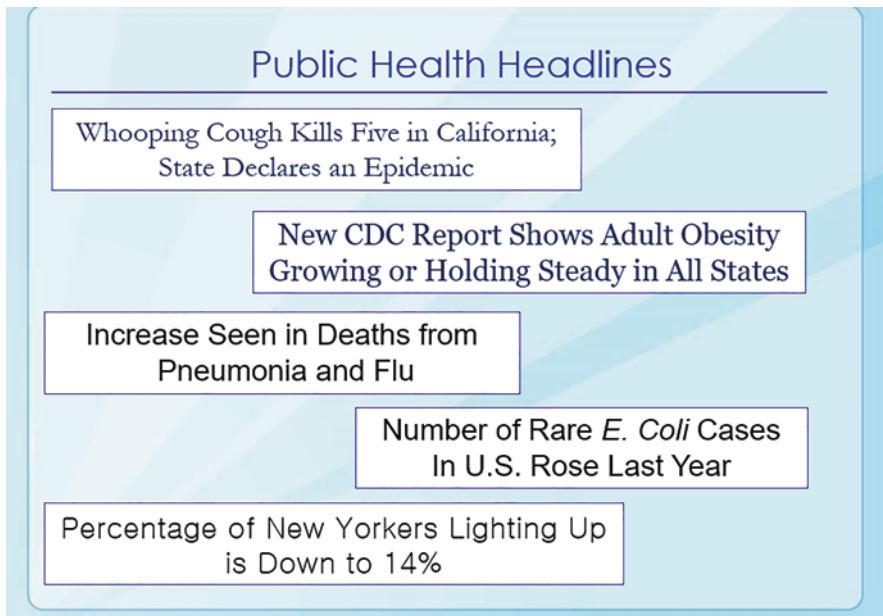
1. Quick detection of epidemics, health problems, and/or changes in health behaviors
2. Estimation of magnitude and scope of the identified health problems
3. Patient observation and contact tracing to determine exposure and treatment options
4. Examine natural history of a disease
5. Prevent the spread of disease
6. Detect disease outbreak or define a problem
7. Analyze trends and characterize disease
8. Monitor changes in infectious agents
9. Detect changes and observe trends in health practices and behaviors
10. Design and implement programs and control measures
11. Evaluate program effectiveness [2, 6] (Fig. 8.1)

Over the past few decades, disease surveillance has evolved into a uniquely distinct discipline of epidemiology. [6]. There are various conditions for which surveillance is used; infectious diseases, chronic diseases (cancer, malnutrition), accidents, and injuries (intentional – homicide and suicide; unintentional – falls) occupational injuries, health effects of toxic exposures, personal health practices and behaviors (smoking, drug/alcohol/substance use, and sexual behavior).

This chapter will discuss public health surveillance as a tool in epidemiological studies for targeting and monitoring public health interventions.

## 2 Approaches to Public Health Surveillance

An approach to surveillance in public health includes coverage, intensity, standardization, analysis and interpretation, dissemination, and evaluation.



**Fig. 8.1** Newsworthy headlines highlighting past epidemics which public health surveillance provided information about. (Source – Centers for Disease Control and Prevention [1])

## 2.1 Coverage

Coverage is the level of inclusion within a specified population. Coverage employs two main strategies; universal or sentinel. Universal coverage chooses a representative sample or an entire population to monitor and observe a disease or condition of interest. Example includes surveillance for food poisoning, bioterrorism agents, and measles. Sentinel coverage on the other hand chooses specific location to observe and monitor a disease or condition of interest. It could also involve choosing a site or location that is most susceptible. For example, selecting a specific provider to monitor a condition among service users, another example is surveillance of animals/vectors.

## 2.2 Intensity

Intensity describes the kind of approach used in public health surveillance. It could be active or passive. A public health surveillance approach is active when public health officials or surveillance investigator goes directly to providers actively find cases and passive when the providers and hospitals are saddled with the responsibility of reporting the cases to surveillance investigators or the health department.

## ***2.3 Standardization***

Standardization an important public health tool used for evaluating population health status by comparing the health of populations after removing the effects of confounding and differences or disparities in population characteristics. It involves calculating rates and adjusting the calculated rates. It allows comparing morbidity and/or mortality rates in two or more different geographic locations even if there are differences in frequency of population distribution. There are two methods, direct and indirect standardization, if age-specific rates for two or more population are known, it is direct standardization. Standardization is dependent on case definition, data collection, and data processing and management [7, 8].

## ***2.4 Analysis and Interpretation***

Analysis and interpretation data analysis factors in person time and place; identifies conditions with higher-than-expected frequency, examines changes over time, and cluster cases. Various analysis methods are used based on the study type. Data analysis uncovers study findings, while data interpretation gives meaning to the data.

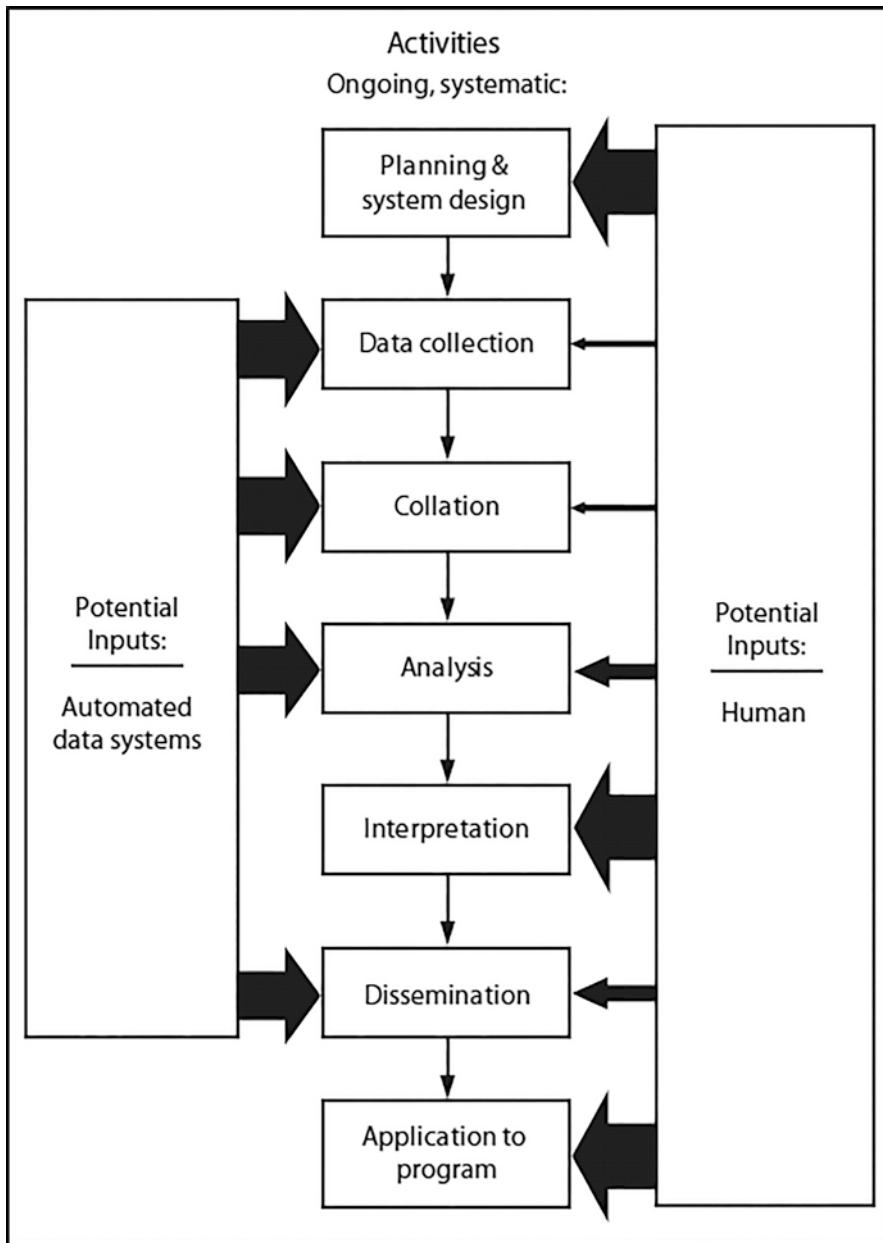
## ***2.5 Dissemination***

Dissemination involves information sharing where data is transmitted to the end users which could be community, consumers, or customers. It involves the publishing or release of data and valuable information derived from a research or statistical activity to users or general public using various media, for example, fact sheet, press release, original article, and news.

## ***2.6 Evaluation***

Evaluation is also an important public health tool used to measure the extent to which a public health intervention was successful. The main purpose of evaluation is to determine intervention effectiveness and to assess and improve the quality of the intervention. Evaluation can be done before, during, and after an intervention or all through the intervention. Formative evaluation is done before or in the early stages of an intervention, process evaluation occurs during the intervention, while summative evaluation takes place after the intervention and evaluates the overall process of the intervention [9].

Figure 8.2 outlines the processes involved in completing a surveillance.



**Fig. 8.2** Processes involved in conducting a surveillance

### 3 Common Terms Used in Public Health Surveillance

**Event** occurrence or development of a condition or an illness, injury in a population which poses immediate threat to human health, thus requiring prompt action. Public health events may be biological, chemical, or radiological in origin, and the severity may differ greatly and require varying degrees of response [10].

**Case** referred to as a person identified to having a particular condition, disease, or health disorder in a study group or population under investigation. A case can be termed “**confirmed**” if there is a laboratory confirmation of disease-causing agent, or “**probable**” if typical clinical features of illness is observed while laboratory result confirmation is pending or there is an epidemiologic link to a laboratory confirmed case, or “**suspected**” if typical clinical features of illness is observed and laboratory or epidemiologic information is missing.

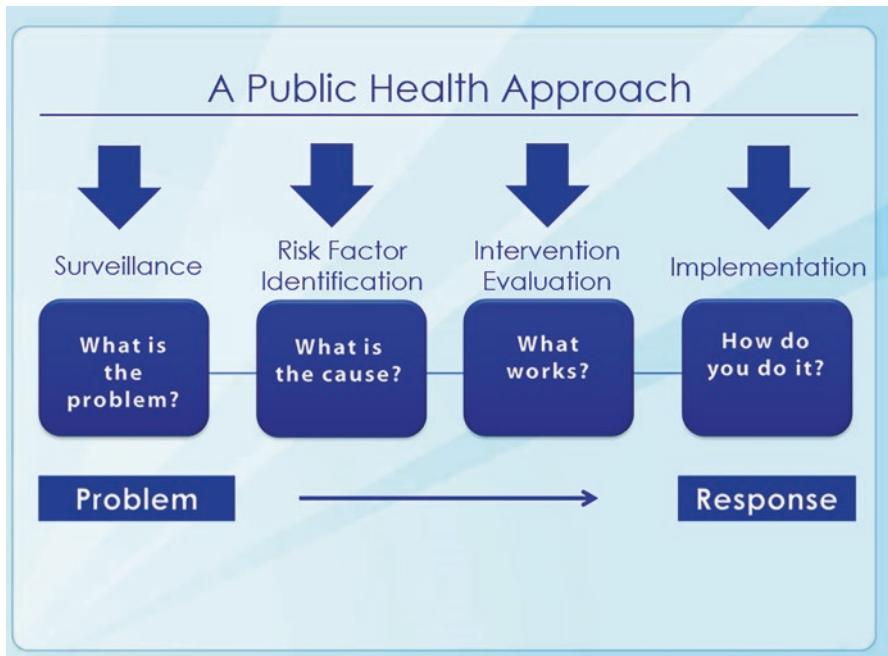
**Rate** is a measure of proportion of occurrence of an event in a defined population over a specific period of time. It is the number of occurrence/cases divided by total population at risk during the specified time period of the occurrence. Number of events is the numerator and total population at risk is the denominator.

$$\text{Rate} = \frac{\text{Number of events in a specified period (numerator)}}{\text{Total population at risk of the events in a specified period (denominator)}}$$

**Examples** Incidence rate measure proportion of new cases (numerator), prevalence rate measure proportion of existing cases (numerator), morbidity rate measure proportion of a disease condition (numerator), and mortality rate measure proportion of deaths (numerator).

**Indicator** a measurement to identify how potent and widespread a hazard is in a population. Indicator is measure of risk posed to human health. Environmental quality of a community for example is a good health indicator of that community. Morbidity indicator also measures the health index of a community.

**Exposure and Outcome** Exposure refers to any measurable factor that may predict the presence of a stated outcome or be associated with a disease/health while an outcome is the actual characteristic or disease condition that is being predicted, for example, the association of smoking (exposure) and increased risk of lung cancer (outcome) (Fig. 8.3, Table 8.1).



**Fig. 8.3** Public health approach to surveillance summarized. (Source – Centers for Disease Control and Prevention [1])

## 4 Types of Public Health Surveillance

According to the CDC, surveillance systems can be broadly categorized into infectious diseases, chronic or noninfectious disease conditions, both infectious and chronic disease and health conditions, and risk factors/exposures. Public health surveillance funding has increased over time [11]. Donors and government agencies at federal, state, and local levels rely on evidence-based data provided by effective surveillance systems. There are various types of surveillance.

### 4.1 Active Surveillance

Active Surveillance is a system of surveillance where data collection from physicians, healthcare providers, laboratories, or directly from the population are initiated by local or state department of health. May be referred to as case finding. In active surveillance, the health department is very proactive about disease investigation and case finding, thus providing a more complete estimate of frequency of

**Table 8.1** Differences between public health surveillance and epidemiologic research

	Public health surveillance	Epidemiologic research
Reason for initiating data collection	Problem detection Problem description Identify cases for epidemiologic studies May be legally required Monitor geographical and temporal trends in disease occurrence	Hypothesis testing Problem description
Frequency of data collection	Ongoing and continuous	Usually one time and time limited
Method of data collection	Established systems or procedures Many persons involved Traditionally depends on voluntary participation	Special procedures tailored to hypothesis or question of interest Fewer persons involved Depends on paid, supervised employees
Amount of data collected per case	Usually minimal	Can be considerable and usually detailed
Completeness of data collected	Often incomplete	Usually complete
Analysis of data	Traditionally simple Primarily to detect change in incidence Usually historical comparison groups	Can be complex Hypothesis testing often requires statistical methods Concurrent controls
Dissemination of data	Timely Regular Review in public health agency Targeted to public health and clinical audience	Not timely Sporadic External review Targeted to academic as well as public health and clinical audience
Use of data	Identifies a problem Triggers intervention Suggests hypotheses Commonly used to evaluate programs Estimates magnitude of a problem	Describes a problem in detail Provides etiologic information Tests hypotheses, suggests additional hypotheses Less often used to evaluate programs

Source: Thacker and Berkelman [2]

disease. Staff members are employed to regularly contact the above listed to seek information about health conditions. Active surveillance is very useful during outbreak investigation because it provides complete, timely and accurate information; however, it is expensive and time consuming. An example is the Behavioral Risk Factor Surveillance System (BRFSS). Active surveillance has been employed in tracking the progression of different types of cancers [12] like prostate, thyroid, renal cancer, and some emerging diseases during outbreak investigation.

## 4.2 Passive Surveillance

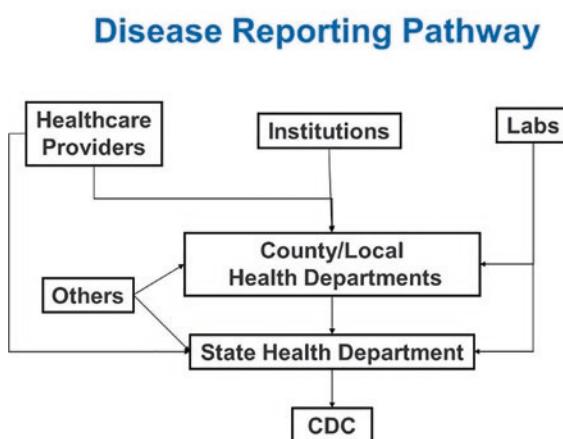
Passive surveillance is the commonest form of surveillance where physicians, healthcare providers, hospitals, clinics, and laboratories regularly report cases to local health departments. May be referred to as case reporting. The reports submitted basically follows procedure of standard case definition of a disease condition. Although passive surveillance is relatively cheap and cover a large area/population, data quality, and timeliness of report cannot be ascertained, also there is no practical way to ensure and enforce providers' adherence to case reporting. A good example is reporting notifiable diseases and conditions. This method of surveillance is very useful in identifying outbreaks and trends over a period.

## 4.3 Syndromic Surveillance

Syndromic surveillance is a relatively new surveillance system which uses case definitions that are based solely on visible symptoms or clinical features without any diagnostic test. The premise of syndromic surveillance is that a syndrome is a combination of signs and symptoms. It is commonly used in resource limited settings but lacks specificity. This is because different disease conditions may exhibit the same symptoms and cause misdiagnoses. An example is collecting data on diarrhea cases rather than cholera cases because not all diarrhea cases are cholera [13, 14].

Figure 8.4 depicts disease reporting pathway adopted by most healthcare providers and health departments across the United States.

**Fig. 8.4** Disease reporting system for surveillance programs



#### ***4.4 Integrated Surveillance***

Integrated surveillance is a combination of both active and passive surveillance which uses the same single infrastructure to collect and collate data about a multiple diseases or behaviors. This surveillance system may be less effective and may lead to data duplication. An example is collecting data on causes of illnesses or deaths.

#### ***4.5 Sentinel Surveillance***

Sentinel surveillance monitors the frequency of specific conditions to assess changes in health status or stability of a population. Sentinel surveillance identifies all cases in a specific location/site based on laboratory confirmation to characterize risk factors and trends associated with a disease. This type of surveillance focuses on a single site or few health facilities which are responsible for collecting data on cases with specific case definitions under surveillance like influenza or diarrhea. Most sentinel sites do not have target population or denominator to calculate incidence, therefore only number of cases are recorded [15].

### **5 Process of Public Health Surveillance**

An effective surveillance system not only collects, analyzes, and disseminates data to policy makers and public health leaders, but it should also address specific health outcomes efficiently and protect population health.

Listed below are some of the key features of an ideal surveillance system:

- (i) Simple: A surveillance system must be easy to use and understand, daily operations, and technical know-how is not cumbersome.
- (ii) Timely: must be able to provide needed information quickly. Data that is unavailable when needed is not reliable for program planning. Report of an infectious disease for example must be very fast to prevent spread of disease.
- (iii) Flexible: surveillance system must be able to accommodate changes and adapt to different situations and respond to new demands, that is, must work effectively during outbreak situations or when a disease condition is endemic.
- (iv) Representative: survey or data collected must truly be reflective of the population of interest. Information collected must be generalizable.
- (v) Generally acceptable: a good surveillance system must be readily acceptable both to the public and healthcare providers. This reflects the willingness of organizations and individuals to participate in the system and its processes.
- (vi) Sensitive: few or no missing information/cases, the surveillance system must be able to specifically identify and detect the health event it was designed to.

- (vii) Strong predictive value: very useful in identifying rare notifiable diseases, involves adopting restrictive case definition so there are no testing errors or false positive and false negatives.
- (viii) Cost effective: inexpensive and relatively cheap [5, 16].

## 6 Public Health Surveillance System

is referred to as a collection of components and processes which enable public health professionals successfully conduct and complete surveillance investigation.

Enabling components of surveillance system include:

1. Laboratory diagnostics
2. Information technologies
3. Clinician consultation
4. Case reporting
5. Public health workforce (clinicians, public health workers, laboratory workers, educators, and communicators)
6. Legislations
7. Policies
8. Regulations
9. Administrators

The goal of an effective surveillance system is to provide information needed to guide public health decisions across different areas ranging from resource allocation, disease prevention, health promotion, prevention program planning and evaluation, and quality improvement [2, 4].

A surveillance system may use one or combination of different methods and data sources, although reliable sources which collects the most accurate information is usually recommended. In some cases, other data sources could complement the primary data source. The specific methods and sources adopted by a surveillance system is determined by the disease condition being surveilled, method used for case identification and definition, the goal of the surveillance system, availability of personnel and material resources, the target population, and the characteristics of the disease's occurrence. The process of surveillance delineates the stepwise actions taken in planning surveillance. Establishing and maintaining a surveillance system requires following these elements:

1. Establish goals: Goals setting in establishing an effective public health surveillance system is key. Without goals, making decisions about priorities and method become difficult. If a surveillance system must achieve its purpose, simple, measurable, achievable, relevant, and time-bound (SMART) goals must be set.
2. Develop case definition: In public health, case definition answers the question of who or what counts as a case and describes the target population in measurable terms. Case definition is very specific with time, place, person, features/behav-

ior, exposure, and diagnosis. When developing case definition, there should be a balance between reliability and simplicity of application within a surveillance system.

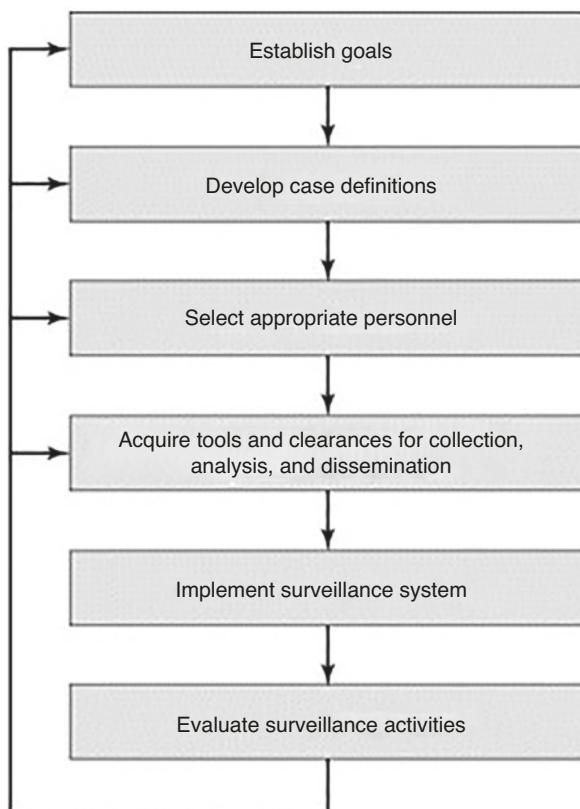
3. Select appropriate personnel: Surveillance investigators need adequate training on outbreak investigation and technical know-how of the surveillance system being established. A public health personnel must be able to correctly navigate a surveillance system for data entry, analysis and interpretation as well as for data sharing and information dissemination.
4. Acquired needed tools for data collection, analysis, and dissemination: The appropriate system infrastructure, hardware, software application, and tools employed while developing and implementing a public health surveillance system is very key. A cloud-based server potentially provides a more cost-effective solution and enhances data access and sharing across different levels of the health system.
5. Implement surveillance system: involves establishing a surveillance database which has the ability to automate data exchange between hospitals/providers information system and health information platform across different administrative levels.
6. Evaluate surveillance activities: evaluating public health surveillance systems helps understand system performance and provides guidance for improvement. Routine evaluation of surveillance activities is recommended and can be done at any point to ensure set goals are being met and to assess effectiveness [10, 17–20]. Ideally, every public health programs should meet stated goals and objectives.

Figure 8.5 highlights the elements of a public health surveillance system.

In conclusion, disease surveillance in local setting is the bedrock of public health tracking and prevention. Public health surveillance helps identify disease trends and emerging public health concerns, also identifies potential points of intervention. Also, surveillance data can provide guidelines for evaluating intervention measures for preventing spread of disease in populations and enable health experts set priorities and policies [21].

### Case Study 1

The year is 2023, and Johnathan walks into a UPS store in Windsor, Ontario, Canada. Johnathan walked to the counter and handed the employees a box. The box was filled with snacks he gathered to send to some friends in Tapa Munto, to lift his spirits. Johnathan had recently returned from there and knew his friend had been going through a rough patch and had recently lost some family members and friends. Johnathan handed the box to the UPS worker and suddenly felt a bit dizzy. He took a moment to compose himself before opening his eyes and smiling at the worried looking clerk. The worker gestured to a seat, and Johnathan sat down heavily, while explaining that he traveled from Hong Kong, and had only been in the country for a few days. He was tired, sore, and had skipped a few meals due to traveling. Thankfully the store wasn't busy, so the clerk was able to process his order quickly. The employee handed Johnathan a receipt, and he took it, thanking her as he winced,



**Fig. 8.5** A public health surveillance system

and headed toward the door. Before he could make it, he collapsed, and began vomiting, while clutching his stomach. The employee called 911 anxiously while glancing in Johnathan's direction. She tells the operator that a customer came in and collapsed. The 911 operator sends an ambulance, and in the meantime tries to get more information from the employee. The operator confirms Johnathan's status with the worker.

According to the employee, Johnathan appeared to be breathing but was not conscious. The employee also told the operator Johnathan looked tired, and she was trying to quickly process his package so he could go home. Two EMT's walk through the door, with a gurney and immediately head to Johnathan, as the employee updates the operator, and hangs up the phone. As the EMT's begin tending to the patient, one of them notices the vomit, and glances at his coworker, who was currently taking Johnathan's pulse. The EMT's also notice he has a bruise on his forearm and appeared to be sweating. The EMT'S begin loading Johnathan onto the gurney after several tries of trying to wake him. The employee begins to nervously ramble about Johnathan's recent trip, and abroad, and if he could just be fatigued, or

have a tape worm. One of the EMT's stop and ask the employee if she knew where Johnathan had traveled. She informed the employee of his trip to Hong Kong, and about his package he sent to his friend, who had recently had a few family deaths. The EMT rushed toward the door, and locked it as he pulled out his phone, as he remembered hearing about an EBOLA outbreak in Asia. He turned to his coworker, "Put on some gloves!" He glanced at the UPS worker, "Is there anyone else in here?" The UPS worker nodded, and confirmed she had a coworker organizing packages in the back.

**Study Problem** If you were hired as a contract Surveillance Epidemiologist and this case was reported in your district, how would you handle this case?

### Case Study 2

Several years ago, some new investors trooped in to Gustivo, a local community in East Africa to mine coals for export after signing a deal with the community leaders. They hired agile and willing young men within the community as mine workers and who were robustly compensated. These community workers were provided with requisite protective gears for the job but they ignored using it, claiming the gears slowed down their daily work/activity. As time goes by the workers also trained their sons, cousins, and nephews who showed interest in the mining job. Recently, Gustivo community is experiencing a higher than usual deaths especially among men. Prior to deaths, the community health center nurse noted that many patients complained of cough, severe headaches, coughing up blood, difficulty breathing, wheezing, tightness of chest, weight loss, and general body pain. The nurse added that malaria is the usual illness treated in the community and have these kinds of symptoms have never been reported in the health center. These cases were reported to the national ministry of health where you work as an independent consultant.

**Study Problem** What steps would you take first to investigate?

Assuming disease reporting system in the community is non-existent or not fully developed, the national ministry of health wants to establish a public health surveillance system where disease reporting will be prompt and timely.

**Study Problems** How will you advise board of directors? How will you establish a public health surveillance system in this community?

### Further Practice

1. \_\_\_\_\_ is the chief cornerstone of the entire public health practice?
  - (a) Data collection
  - (b) Health planning
  - (c) Decision-making
  - (d) Public health surveillance
  - (e) All of the above

2. The main goal of public health surveillance is to provide valuable information needed for decision making to prevent and control disease.
  - (a) True
  - (b) False
3. All the following are some benefits of public health surveillance, except
  - (a) Examine natural history of a disease
  - (b) Prevent the spread of disease
  - (c) Detect disease outbreak or define a problem
  - (d) Analyze trends and characterize disease
  - (e) Account for government funding on a public health issue
4. Only infectious diseases need surveillance activities.
  - (a) True
  - (b) False
5. Sentinel coverage chooses a representative sample or an entire population to monitor and observe a disease or condition of interest.
  - (a) True
  - (b) False
6. Which of the following is true about “active surveillance”?
  - (a) It is also referred to as case reporting.
  - (b) It is not useful during outbreak investigation.
  - (c) It is initiated by local or state department of health.
  - (d) There is frequent transfer of data on disease from state health departments to CDC.
7. Passive surveillance is a rarely used type of public health surveillance.
  - (a) True
  - (b) False
8. Features of an ideal surveillance system include all but one of the following
  - (a) Representative
  - (b) Sensitive
  - (c) Complex
  - (d) Flexible
  - (e) Timely
9. The major goal of an effective surveillance system is to provide information needed to guide public health decision data interpretation.
  - (a) True
  - (b) False

10. Case definition is very specific with time, place, person, features/behavior, exposure, and diagnosis.
  - (a) True
  - (b) False

## Answer Keys

1. (d)
2. (a)
3. (e)
4. (b)
5. (b)
6. (c)
7. (b)
8. (c)
9. (a)
10. (a)

## References

1. Centers for Disease Control and Prevention (CDC). Introduction to public health. In: Public health 101 series. Atlanta, GA: U.S. Department of Health and Human Services, CDC; 2014. Available at: <https://www.cdc.gov/training/publichealth101/surveillance.html>. Accessed 8 Mar 2023.
2. Thacker SB, Qualters JR, Lee LM, Centers for Disease Control and Prevention. Public health surveillance in the United States: evolution and challenges. MMWR Suppl. 2012;61(3):3–9.
3. Office of Public Health Scientific Services, Centers for Disease Control and Prevention. Public health surveillance: preparing for the future. Atlanta: Centers for Disease Control and Prevention, CDC; 2018. Available at: <https://www.cdc.gov/surveillance/pdfs/Surveillance-Series-Booklet.pdf>
4. Groseclose SL, Buckeridge DL. Public health surveillance systems: recent advances in their use and evaluation. Annu Rev Public Health. 2017;38:57–79.
5. Tulchinsky TH. John snow, cholera, the broad street pump; waterborne diseases then and now. In: Case studies in public health [internet]. Elsevier; 2018. p. 77–99. Available from: <https://linkinghub.elsevier.com/retrieve/pii/B9780128045718000172>.
6. Declerck S, Carter AO. Public health surveillance: historical origins, methods and evaluation. Bull World Health Organ. 1994;72(2):285–304.
7. Boslaugh S, editor. Encyclopedia of epidemiology. Sage Publishing; 2007.
8. Naing NN. Easy way to learn standardization: direct and indirect methods. Malays J Med Sci. 2000;7(1):10–5.
9. Smith ML, Ory MG. Measuring success: evaluation article types for the Public Health Education and Promotion Section of Frontiers in Public Health. Front Public Health. 2014;2:111. <https://doi.org/10.3389/fpubh.2014.00111>.
10. Carter A, National Advisory Committee on Epidemiology Subcommittee. Establishing goals, techniques and priorities for national communicable disease surveillance. Can J Infect Dis. 1991;2(1):37–40. <https://doi.org/10.1155/1991/346135>.

11. Office of Public Health Scientific Services, Centers for Disease Control and Prevention. Public health surveillance: preparing for the future. Atlanta: Centers for Disease Control and Prevention; 2018. Available from: Public Health Surveillance Preparing for the Future ([cdc.gov](http://cdc.gov))
12. Carter HB. Optimizing active surveillance. *Eur Urol.* 2016;70(6):909–11. <https://doi.org/10.1016/j.eururo.2016.07.017>.
13. Lyerla R, Stroup DF. Toward a public health surveillance system for behavioral health. *Public Health Rep.* 2018;133(4):360–5.
14. Paterson BJ, Durrheim DN. The remarkable adaptability of syndromic surveillance to meet public health needs. *J Epidemiol Glob Health.* 2013;3(1):41–7. <https://doi.org/10.1016/j.jegh.2012.12.005>.
15. Murray J, Cohen AL. Infectious disease surveillance. In: International encyclopedia of public health; 2017. p. 222–9. <https://doi.org/10.1016/B978-0-12-803678-5.00517-8>.
16. Lucero-Obusan C, Oda G, Mostaghimi A, Schirmer P, Holodniy M. Public health surveillance in the US Department of Veterans Affairs: evaluation of the Praedico surveillance system. *BMC Public Health.* 2022;22(1):272. <https://doi.org/10.1186/s12889-022-12578-2>.
17. Sheikhali SA, Abdallat M, Mabdalla S, Al Qaseer B, Khorma R, Malik M, Profili MC, Rø G, Haskew J. Design and implementation of a national public health surveillance system in Jordan. *Int J Med Inform.* 2016;88:58–61. <https://doi.org/10.1016/j.ijmedinf.2016.01.003>.
18. Bilandzic A, Bozat-Emre S. At-a-glance-initial evaluation of Manitoba's cannabis surveillance system. *Health Promot Chronic Dis Prev Can.* 2020;40(7–8):245–9. <https://doi.org/10.24095/hpcdp.40.7/8.04>.
19. Burkhardt H, Loschen W, Wojcik R, Holtry R, Punjabi M, Siwek M, Lewis S. Electronic surveillance system for the early notification of community-based epidemics (ESSENCE): overview, components, and public health applications. *JMIR Public Health Surveill.* 2021;7(6):e26303. <https://doi.org/10.2196/26303>.
20. Jiang WX, Huang F, Tang SL, Wang N, Du X, Zhang H, Zhao YL. Implementing a new tuberculosis surveillance system in Zhejiang, Jilin and Ningxia: improvements, challenges and implications for China's National Health Information System. *Infect Dis Poverty.* 2021;10:22. <https://doi.org/10.1186/s40249-021-00811-w>.
21. Aiello AE, Renson A, Zivich P. Social media-and internet-based disease surveillance for public health. *Annu Rev Public Health.* 2020;41:101–18.

# Chapter 9

## Standardization



Anwar T. Merchant

### Learning Objectives

- Describe reasons for using standardization in epidemiology
- Calculate direct standardization
- Calculate indirect standardization
- Illustrate the methods of standardization with examples

## 1 Introduction

Suppose we want to find out if people living in community A are healthier than those living in community B. Both communities are approximately the same size. To do that, we compare the number of people dying yearly in communities A and B. We find that there were more deaths in community B than in A. However, one analyst notices that people living in community A are mostly young, and those in community B are primarily old. Based on these data, can we say that people living in community A are healthier than those in community B? Not necessarily, because the risk of death increases with age, and communities A and B differ. We are comparing apples and oranges. We want to ask, “Would there be more deaths in community B if the ages of people living there were similar to those living in community A?”

---

A. T. Merchant (✉)

Epidemiology Division, Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA  
e-mail: [merchant@mailbox.sc.edu](mailto:merchant@mailbox.sc.edu)

## 2 Standardization

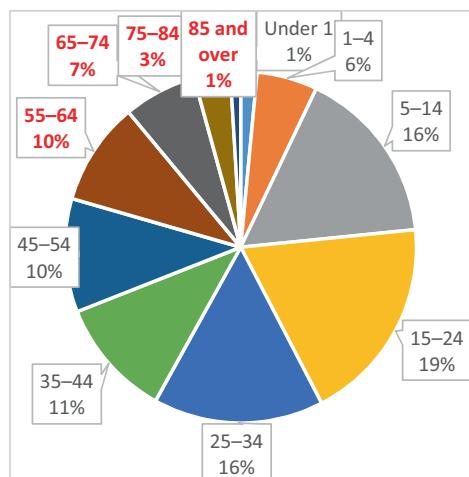
Standardization is done to make data comparable. For example, the USA's crude death rates per 100,000 in 2019 and 1978 were 870 and 868, respectively [1, 2]. This means that approximately the same proportion of people died in the USA each year in 2019 and 1978. A naïve comparison of crude death rates suggests that changes in living conditions, smoking habits, diet, and health care between 1978 and 2019 have had no effect on death rates. However, many things that affect death rates have changed between those years, one of the most important being age. The population age distributions in 1978 and 2019 are shown in Figs. 9.1 and 9.2, respectively. There are more older adults in the USA in 2019 compared with 1978, as seen in Figs. 9.1 and 9.2. For example, 13% of the population was between 55 and 64 years old in 2019 compared to 10% in 1978. A similar trend is seen in the other older age groups as well. As the risk of death increases with age, we expect more deaths in 2019 due to old age. Standardization makes it possible to compare the death rates in 1978 and 2019 after controlling for age.

## 3 Direct Standardization

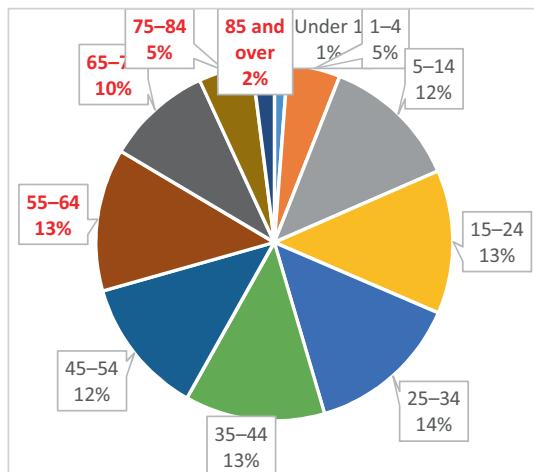
This is one method to compare death rates in 1978 and 2019 controlling for age. The first step is to get the distribution of deaths in 2019 by age group. These data are used to calculate crude and age group-specific death rates. This is shown in Table 9.1.

The crude death rate per 100,000 population is calculated by dividing the number of deaths in 2019 by the total population in 2019 and multiplying by 100,000. This is the number of people dying per 100,000 population in 2019, called the crude death rate. The formulas are given below.

**Fig. 9.1** Age distribution of US population in 1978 [1]



**Fig. 9.2** Age distribution of US population in 2019 [2]



**Table 9.1** Deaths by age group in the USA in 2019

	US deaths 2019	US population 2019	Deaths per 100,000 in 2019
Age in years			
Under 1	20,921	3,915,337	534
1–4	3676	15,661,346	23
5–14	5497	40,994,163	13
15–24	29,771	42,687,510	70
25–34	59,178	45,940,321	129
35–44	82,986	41,659,144	199
45–54	160,393	40,874,902	392
55–64	374,937	42,448,537	883
65–74	555,559	31,483,433	1765
75–84	688,027	15,969,872	4308
85 and over	873,746	6,604,958	13,229
Total	2,854,691	328,239,523	

$$\text{Crude death rate} = 100,000 \times \left( \frac{\text{Number of deaths in year}}{\text{Total population in year}} \right)$$

$$\text{Crude death rate} = 100,000 \times \left( \frac{2854691}{328239523} \right) = 870$$

The age-specific death rates per 100,000 population are calculated by dividing the number of deaths in a specific age group in 2019 by the population in that age group in 2019 and multiplying by 100,000. This is the proportion of people in the under 1 year age group dying per 100,000 population in 2019 and is called

age-specific death rate for under 1 year. The calculation of the age-specific death rate for the under 1 year age group is given below.

$$\text{Age specific death rate} = 100,000 \times \left( \frac{\text{Number of deaths in year in age group}}{\text{Population in that age group}} \right)$$

$$\text{Age specific death rate Under 1 year} = 100,000 \times \left( \frac{20921}{3915337} \right) = 534$$

Age-specific death rates for the rest of the age groups in Table 9.1 are calculated similarly. This is the first step in direct standardization.

The next step is to select a standard population. We can choose the 1978 US population as the standard population in this case. By doing this, we will estimate the death rate in 1978 if that population had the age-specific death rates observed in 2019. This is shown in Table 9.2.

To do that, we first need to estimate the number of deaths in each age group in 1978 if the age-specific death rate was observed in 2019. For example, the age-specific death rate for the under-1 group was 534/100,000 in 2019. The number of individuals in that age group in 1978 was 3,326,000. The expected number of deaths in that age group is obtained by multiplying these two numbers, as shown below.

$$\begin{aligned} & \text{Expected deaths in Under 1 group in 1978} \\ &= \text{Age specific death rate} \times \# \text{ in that age group} \end{aligned}$$

$$\text{Expected deaths in Under 1 group in 1978} = \frac{534}{100,000} \times 3,326,000 = 17772$$

**Table 9.2** Direct standardization using US 1978 population as standard

	US deaths 2019	US population 2019	Deaths per 100,000 in 2019	US population 1978	Expected deaths in 1978 population
Age in years					
Under 1	20,921	3,915,337	534	3,326,000	17,772
1–4	3676	15,661,346	23	12,409,000	2913
5–14	5497	40,994,163	13	36,220,000	4857
15–24	29,771	42,687,510	70	42,183,000	29,419
25–34	59,178	45,940,321	129	34,803,000	44,831
35–44	82,986	41,659,144	199	24,373,000	48,552
45–54	160,393	40,874,902	392	23,166,000	90,903
55–64	374,937	42,448,537	883	21,112,000	186,477
65–74	555,559	31,483,433	1765	14,996,000	264,621
75–84	688,027	15,969,872	4308	7,412,000	319,330
85 and over	873,746	6,604,958	13,229	2,095,000	277,140
Total	2,854,691	328,239,523		222,095,000	1,286,814

Expected deaths are estimated in the same way for each of the other age groups. The total number of expected deaths in 1978 is obtained by summing the age-specific deaths. This is the expected number of deaths in 1978 if that population had the 2019 age-specific death rates. In this case, it is 1,286,814. To get the standardized death rate per 100,000 in 1978, the expected deaths in 1978 are divided by the total population in 1978 multiplied by 100,000 as shown below.

Standardized death rate 1978

$$= 100,000 \times ((\text{Expected deaths in 1978}) / (\text{Population 1978}))$$

$$\text{Standardized death rate 1978} = 100,000 \times \left( \frac{1,286,814}{222,095,000} \right) = 579$$

More generally,

$$\text{Standardized death rate} = \left[ \frac{\sum_i a_i \times r_i}{\sum_i a_i} \right] \times 100,000$$

$a_i$  = population in age group  $i$  of standard population

$r_i$  = death rate in age group  $i$  of comparison population

The crude death rate in 2019 was 870 per 100,000 population, but the standardized death rate using the 1978 age distribution as the standard was 579 per 100,000. The comparison of crude rates was biased because the age distributions of the populations in 2019 and 1978 were different. Direct standardization corrected for that bias. The 2019 death rate standardized to the 1978 population is lower than the crude death rate in 1978 (579 versus 868 per 100,000), suggesting that the population in 2019 is living longer than in 1978.

#### Box 1 Summary of Steps for Direct Standardization [3]

- Get distribution of population by age (2019 data)
- Get distribution of deaths by age (2019 data)
- Calculate the rate of death by age category
- Identify the standard population (1978 data)
- Get distribution of standard population by age
- Calculate expected number of death for each age group of the 1978 population using 2019 age-specific death rates
- Sum up the expected deaths in the 1978 population and divide by 1978 total population to get the age-standardized death rate
- Note: The data could be standardized factors other than age if that factor predicted the death (e.g., sex)

## 4 Indirect Standardization

In the previous example, we had mortality data from the entire US population in 2019. Frequently, such a large amount of data is not available. Consider this hypothetical example. The Mayor suspects more people in his town, a population of 7648, are getting COVID-19 than elsewhere and asks the local public health officer's opinion. She asks the Mayor why he suspects that. He says that in the last month, 252 people, including several people he knew, had COVID-19.

The number of cases and the town's population is too small to estimate a rate accurately. The public health officer decides to use indirect standardization.

- First, she finds out from the Johns Hopkins Coronavirus Resource Center website that there were 144,793 new COVID-19 cases in the last month in the state, population 11 million.
- She calculates the COVID-19 infection rate in the last month in her state as shown below.

- Rate in state =  $\left( \frac{\# \text{ new COVID cases}}{\text{Total population}} \right) \times 100,000$
- Rate in state =  $\left( \frac{144,793}{11,000,000} \right) \times 100,000 = 1,316$
- She then calculates the expected number of COVID-19 cases in her town by multiplying the rate in the state by the town's population as shown below.

- Expected cases =  $\frac{1316 \times 7648}{100,000} = 101$
- Now she divides the observed COVID-19 cases in the last month by the expected number to get a standardized morbidity ratio (SMR) as shown below.
- SMR =  $\left( \frac{\text{Observed}}{\text{Expected}} \right) = \left( \frac{252}{101} \right) = 2.5$
- She explains to the Mayor that his guess about more COVID-19 in the town was probably correct. People in the town are 2.5 times more likely to get COVID-19 than people in the rest of the state.

Note that if the SMR = 1, it implies that the observed and expected deaths are very similar and there is no difference. If the SMR > 1, it implies more deaths than expected, and if the SMR < 1, then there are fewer deaths than expected in the study population.

**Box 2 Summary of Steps for Indirect Standardization [3]**

- Count the observed deaths over a defined period
- Get the total population giving rise to the deaths
- Obtain a death rate from a source that is:
  - Comparable (similar to the observed group)
  - Reliable (accurate)
- Calculate the expected deaths using the death rate
- Divide observed by expected deaths to get SMR
- Interpretation of SMR
  - $SMR = 1 \Rightarrow$  no difference
  - $SMR > 1 \Rightarrow$  more deaths than expected
  - $SMR < 1 \Rightarrow$  fewer deaths than expected

**5 Further Practice**

1. Data from Canada in 2019 is given below.

Population of Canada			
Age group	2019		
	Persons	Death rate per 100,000	Number of deaths
All ages	37,601,230	7.6	
0 to 4 years	1,932,784	1.04	
5 to 9 years	2,041,278	0.1	
10 to 14 years	2,033,308	0.1	
15 to 19 years	2,114,650	0.3	
20 to 24 years	2,475,503	0.5	
25 to 29 years	2,626,204	0.7	
30 to 34 years	2,605,394	0.8	
35 to 39 years	2,581,046	1	
40 to 44 years	2,421,889	1.2	
45 to 49 years	2,398,378	1.8	
50 to 54 years	2,505,044	2.9	
55 to 59 years	2,751,672	4.5	
60 to 64 years	2,514,070	7	
65 to 69 years	2,098,142	10.9	
70 to 74 years	1,708,613	17.1	
75 to 79 years	1,165,334	28	
80 to 84 years	789,039	49.2	
85 to 89 years	513,205	88.7	
90 + years	325,677	188.1	

- (a) Calculate the number of deaths in each age category.
- (b) Calculate the age-adjusted death rate for 2020 using 2019 data. (Assume that the age distribution of the population in 2019 and 2020 was similar.)
- (c) The crude death rate in 2020 in Canada was 8.1 per 100,000. Compare this with the age-adjusted death rate for 2020. Explain why the rates are similar or different.
2. The table below describes the population of a hypothetical town and the number of deaths in 2019.

Age distribution	Population	Deaths in 2019
Under 1	1193	10
1–4	4771	36
5–14	12,489	69
15–24	13,005	77
25–34	13,996	108
35–44	12,692	99
45–54	12,453	97
55–64	12,932	81
65–74	9592	92
75–84	4865	39
85 and over	2012	25

- (a) Were there more deaths in this town than in the rest of the USA in 2019? Please explain the reasons for your conclusion. (Use data from Table 9.2 in this chapter.)
- (b) Why did you choose the methods that you did to make this comparison?

## References

- National Center for Health Statistics. Vital statistics of the United States, 1978. Washington; 1982. Report No.: PHS. p. 83–1101.
- Xu J, Murphy SL, Kochanek KD, Arias E. Deaths: final data for 2019. Natl Vital Stat Rep. 2021;70(8)
- Naing NN. Easy way to learn standardization : direct and indirect methods. Malays J Med Sci. 2000;7(1):10–15. PMID: 22844209; PMCID: PMC3406211.

# Chapter 10

## Causal Association



Amal K. Mitra

### Learning Objectives

After completing this chapter, you will be able to:

- Understand relationship between exposure and risk factors with a disease
- Evaluate Bradford Hill's criteria of causation
- Control confounding variables
- Describe Rothman's causal pie with practical examples
- Apply the concept of cause-and-effect relationship in research findings

## 1 Introduction

In epidemiological studies, one of the key issues is inferring whether an association between the observed risk factors and the disease is causal or not. For example, we are exposed to a number of environmental factors such as overcrowding, noise, heat, dust, smoke, chemicals, etc. which may be considered risk factors for asthma. Suppose, in a cross-sectional study, you found a statistically significant association between an air pollutant, that is, fine particulate matter, called PM 2.5 and patients with stage 2 asthma, which is moderate persistent asthma with symptoms occurring on a daily basis. You have collected information on possible other environmental factors, personal habits of smoking and drinking, and food habits. You wanted to establish the causal relationship, if any, between PM 2.5 and asthma. First of all, you have conducted a cross-sectional study. In several chapters of this book (Chaps. 4, 5, and 6), you have learned about epidemiological study designs including

---

A. K. Mitra (✉)

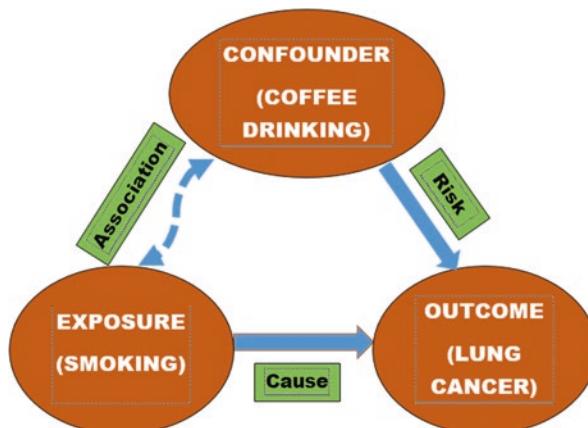
Department of Epidemiology and Biostatistics, Jackson State University, Jackson, MS, USA  
e-mail: [amal.k.mitra@jsums.edu](mailto:amal.k.mitra@jsums.edu)

cross-sectional study, case-control study, and cohort study, where you were informed that certain type of epidemiological studies such as **cross-sectional study** **cannot establish a causal relationship** because temporality cannot be measured in such studies. This chapter will provide you detailed information on several criteria, known as **Bradford Hill's criteria** which will help in assessing whether an association is causal or not. Second, in Chap. 13 of this book, we discussed about **confounding**, or a nuisance factor, which can affect your results. Also, the author discussed the possible methods to control for the confounders. Further to enhance your knowledge, there is a section in this chapter which discusses **Rothman's causal pie** which is used in a situation where there are multiple causal factors for a disease, such as asthma.

## 2 Confounding Effect

Before we talk about establishing a causal relationship, we should make sure that we have taken care of any confounders which can affect the causal pathway between a causal factor and the disease. A *confounder* is a variable which is associated with the exposure factor and the outcome or the disease. A confounding variable is a kind of nuisance variable, which can obliterate the actual finding of data. A confounder competes with the real factor that causes the disease.

Figure 10.1 illustrates the association of two variables smoking and coffee drinking with lung cancer. In this figure, smoking is a cause of lung cancer. Coffee drinking is a confounder which is associated with both the causative factor “smoking” and the outcome of lung cancer. Coffee drinking is one of the risk factors of lung cancer, which affects the actual causal relationship between smoking and lung cancer. Coffee drinking must be controlled to find the real effect of smoking. Let us take



**Fig. 10.1** A schematic diagram showing the effect of a confounder

the example of heart diseases as an outcome variable and physical exercise as the exposure variable. Some of potential confounders that should be taken into account for this kind of study are as follows:

**Age.** Like many other diseases, age is a potential confounder for heart disease. If you want to study the effect of physical exercise on heart disease, you should consider finding after controlling for age. As you know heart disease increases with age, at the same time, older people are more sedentary and likely to be less active or inactive. In this case, age is a potential confounder because it affects both the exposure factor (exercise) and the outcome variable (heart disease).

**Gender.** For most of the diseases, gender is a potential confounder because it can affect both the outcome (disease of interest) and the exposure or the risk factor. Females of reproductive age are less likely to suffer from heart disease compared to males. However, after the natural menopause age of 50 years, the risk of heart disease increases dramatically [1, 2]. The habit of physical exercise may also depend on gender.

**Race.** For the example of heart disease, race is another risk factor. In the data analysis using multivariate regression, you should add physical exercise as well as all other potential risk factors including race in the model.

**Personal habits.** For some diseases, such as heart disease or stroke, the history of personal habits (smoking and drinking) must be considered in the data analysis. If your primary objective is to find the effect of physical exercise on heart diseases, you need to control for smoking and drinking because they are also risk factors for heart disease.

### 3 Hill's Criteria of Causality

An English epidemiologist Sir Austin Bradford Hill suggested nine criteria for assessing causality in 1965 [3]. The more of these criteria are fulfilled; the more likely the association between an exposure and a disease is likely to be causal. These criteria are as follows:

#### 3.1 Strength of Association

Bradford Hill's study of the association between cigarette smoking and lung cancer showed that smokers die 9–10 times more often from lung cancer compared with nonsmokers, and heavy cigarette die 20–30 times more commonly from lung cancer than nonsmokers. Whereas, in the case of coronary thrombosis, the death rates were twice or even less among smokers compared to nonsmokers. He cited another example of the death rate of chimney sweeps from scrotal cancer in the second decade of the twentieth century—it was about 200 times more likely among sweep

workers due to the exposure of tar and mineral oil than who were not exposed to these chemicals [4]. In a prospective cohort study, one of the epidemiological measurements is relative risk. Therefore, the strength of association is more important than statistically significant results to suggest a causal association.

### 3.1.1 Evidence from Longitudinal Follow-Up Studies

#### Example 1 Relationship Between Obesity and Knee Replacement

Among the epidemiological studies, prospective cohort study is one of the study designs that is appropriate for establishing a causal association between an exposure and a disease. In prospective cohort studies, some of the epidemiological measurements include incidence, relative risk (RR), hazard ratio, and attributable risk. A hazard ratio (HR) of more than one indicates an increased risk, and an HR value being less than one indicates a protective effect.

Leyland and colleagues (2016) [5] studied the risk of obesity on the progression of knee osteoarthritis from diagnosis to knee replacement surgery. In a prospective cohort study conducted in a community setting in Spain, they followed up 105,189 participants for a median of 2.6 years (interquartile range, 1.3–4.2 years). Of these patients, 7512 (7.1%) underwent knee replacement. Adjusted hazard ratios and 95% confidence intervals for knee replacement for BMI categories were 1.41 (95% CI 1.27–1.57) for overweight (BMI 25.0–29.9), 1.97 (95% CI 1.78–2.18) for obese-I (BMI 30.0–34.9), 2.39 (95% CI 2.15–2.67) for obese-II (BMI 35.0–39.9), and 2.67 (95% CI 2.34–3.04) for obese-III (BMI  $\geq 40$ ) compared to normal weight (BMI 18.5–24.9), indicating that the risk of knee replacement increased with the increased severity of obesity. Overweight and obese patients were at more than 40% and 100% increased risk of knee replacement surgery, respectively, compared to patients with normal weight.

#### Example 2 Relationship Between Obstructive Sleep Apnea, Diabetes, and Wake-Up Stroke

Wake-up stroke (WUS) are strokes that are noted upon awakening in patients previously going to bed in a normal state of health. Barreto and colleagues (2020) [6] studied 102 adults having obstructive sleep apnea (OSA), who were followed up for 12 months. In addition to classifying the study patients according to severity of sleep apnea (measured by apnea–hypopnea index, AHI), the presence of comorbidities was also recorded. The majority had hypertension (73%), followed by Type 2 diabetes (29.4%), and heart disease (16.7%). A large proportion was physically inactive (69.6%), one-third was smoking (32.4%), and alcohol consumption was among 18%. Wake-up stroke occurred in approximately 1 of 3 cases during the 12-month follow-up period. Interestingly, cases with and without WUS did not differ regarding polygraphic findings of apnea severity. However, the presence of Type 2 diabetes independently increased the risk of WUS by about three times (OR = 2.76; 95% CI 1.10–6.05;  $P = 0.03$ ), compared to those who did not have diabetes.

### 3.1.2 Evidence from Case-Control Study

A review of existing data on dysentery-related deaths in rural Bangladesh suggests that deaths in children followed a recurring seasonal pattern with an increase during the post-monsoon season of August–November of each year [7]. This seasonal pattern of death was not evident among adults. The overall dysenteric death rate was 13.3 per 10,000 population per year. Deaths reported recently by the health workers were re-investigated. Although the causal agents producing fatal dysentery in most patients in the community remained unidentified, it was likely to be species of *Shigella* in childhood deaths. To identify clinical determinants of a fatal outcome, a case–control analysis was done with patients hospitalized with dysentery. Cases were those who died in the hospital due to dysentery. Controls were age-matched survivors of dysentery. The risk factors shown to be significantly associated with children who died, compared to those who survived, were: longer median duration of illness (7 days vs. 2 days;  $p \leq 0.001$ ), female sex (65% vs. 35%; odds ratio [OR] = 4.3; 95% CI = 1.07–18.09;  $p = 0.039$ ), signs of lower respiratory infection (70% vs. 9%; OR = 24.0; 95% CI = 3.71–117.22;  $p \leq 0.001$ ), and severe malnutrition (65% vs. 17%; OR = 8.9; 95% CI = 1.90–45.86;  $p = 0.002$ ) [7].

### 3.1.3 Evidence from Cross-Sectional Study

In a community-based cross-sectional study of 200 people in Odisha, India, Chowdhury et al. (2019) [8] identified risk factors for obstructive sleep apnea. Independent risk factors associated with such subjects were obesity (aOR = 3.5; 95% CI 1.2–10.5), regular alcohol consumption (aOR = 4.5; 95% CI 1.8–11.1), and high blood pressure (aOR = 11.5; 95% CI 4.7–28.0).

All these evidence showed a stronger association between the exposure and the disease outcome, which may suggest a causal relationship. However, cross-sectional studies being descriptive in nature and because of lack of **temporality** of the association in such studies, only risk factors can be identified but **a causal relationship cannot be established** in cross-sectional studies.

## 3.2 Temporality

For the web of causality, a temporal relationship must be established. It is important to remember that the cause of the disease comes first and then the disease. In most of the situations, this is easy to find between two events which comes first and which one follows it. For example, you get an infection first and then get the disease. The infection is the cause of the disease. For some cases, it may be difficult to ascertain this temporal relationship. Let us take an example of malnutrition and diarrhea. We often find that malnutrition and a disease in children such as diarrhea coexist. The question is which comes first? Does a malnourished child often gets

diarrhea—meaning malnutrition is the cause of diarrhea? On the other hand, a child gets diarrhea, the child cannot eat well during diarrhea, or the mother fails to give proper food during diarrhea, and frequent episodes of diarrhea lead to malnutrition. In the latter case, diarrhea (or frequent infections) may be the cause of the child's malnutrition.

### Box 1

In some type of studies, such as in a cross-sectional study, information about both exposure and the outcome (or disease) is obtained at the same time. For example, you are proposing a community-based cross-sectional study for hepatitis B infection. You are gathering information about any existing cases of hepatitis B by doing a door-to-door survey and using a questionnaire. You are also collecting data on what could be cause of hepatitis B. You know that hepatitis B is transmitted primarily from one person to another through sexual contact (semen or vaginal secretion), blood or blood products, and other body fluids (such as saliva). You collected information about possible sources such as sexual contacts, drug habits including sharing needles and syringes, sharing intravenous drugs, and history of blood and blood product transfusion. After data collection and data analyses, you found out some of the risk factors. But how sure are that some of the risk factors are the causes of hepatitis B? In a cross-sectional study, you at best get risk factors but cannot establish a causal relationship between the exposure and the disease because you are not certain which came first.

On the other hand, for the same hypothesis, you may want to conduct a better study design such as **prospective cohort study** with a group of polygamous people with history of exposure to at least one infected partner and a group of monogamous people with no history of exposure, follow the two groups (exposed and nonexposed) for a reasonable time period, and look for the incidence of hepatitis B in both the groups. A prospective cohort study design is a better option to establish causality compared to a cross-sectional study because of having a temporal factor between the exposure and the hepatitis B incidence in a cohort study.

### 3.3 *Consistency*

The term “consistency” is applied when repeated studies show the same result. To confirm the findings of one study, sometime additional studies are undertaken in a different population, or populations of different countries, or studies conducted at different time periods. If several studies point toward similar outcome, the studies are called consistent.

### 3.3.1 Example of Two Clinical Trials Having Consistency

Mitra et al. (1995) [9] evaluated the therapeutic efficacy and safety of an immunological treatment option in reducing diarrhea due to rotavirus in children, using a double-blind controlled clinical trial. The treatment group received a preparation of *hyperimmune bovine colostrum* (HBC) collected from immunized cows with four serotypes human rotavirus. The controls received colostrum from unimmunized cows. After treatment for 3 days, patients who received HBC had a significantly shorter duration of diarrhea than the controls (median 56 hours versus 72 hours; 95% CI = 8–32 hours,  $p < 0.001$ ). In 50% of the children in the study group, diarrhea stopped by 48 h, whereas nearly 100% of the controls was still suffering from diarrhea. Treatment with HBC was found to reduce diarrhea due to rotavirus in children.

In a separate study to children with rotavirus diarrhea, HBC was found effective in reducing significantly daily and total stool output than did children who received placebo ( $P < 0.05$ ). Clearance of rotavirus from the stool was also earlier in the HBC group compared with the placebo group (1.5 days versus 2.9 days,  $P < 0.001$ ) [10]. The two study results were **consistent** in the treatment outcome of rotavirus diarrhea with HBC.

### 3.4 Biological Gradient or Dose-Response

This is one of the commonly used criteria, especially used in the fields of toxicology and pharmacokinetics of drugs. These criteria are also used widely in epidemiology. According to dose-response, an increase (or decrease) in the level, intensity, duration, or total level of exposure to a substance leads to progressive increase (or decrease) in the risk.

**Example 1:** In a hospital-based study in the southern Bangladesh, Mitra et al. (2002) [11] found that increasing doses of arsenic exposure leads to increasing incidence of skin pigmentation and keratosis.

**Example 2:** In Mississippi, a statistically significant association was observed between breast cancer incidence and the total amount of environmental chemicals ( $p = 0.032$ ). Also, more amount of maximum air pollutant emissions was linearly correlated ( $r = 0.24$ ) with more breast cancer incidence in the state [12].

### 3.5 Specificity

Specificity results in when the causal factor (or exposure) leads to only one disease and the disease results from only the single cause. Let us take two examples:

**Example 1:** Lung cancer is developed after cigarette smoking. The question is—does cigarette smoking cause only lung cancer? The answer is—no, cigarette smoking causes lung cancer as well as many other cancers such as cancer of larynx, esophagus, mouth, bladder, liver, stomach, pancreas, colon, etc. Therefore, smoking cigarettes is not specific to lung cancer.

**Example 2:** *Vibrio cholera* bacteria causes cholera. The association is specific because *Vibrio cholera* only causes cholera, and cholera is caused by a specific group of bacteria called *Vibrio cholerae*.

### 3.6 Plausibility

The association should be plausible (or explainable) in terms of known biological facts about the pathogenesis of the disease.

**Example 1:** Many people die in bed. Is it plausible that sleeping in bed causes death? Certainly, it is not the case.

**Example 2:** In an African country, many people are dying from malaria. Suppose, a common drug for the treatment of malaria known as Chloroquine is not working in most patients. Laboratory data shows that the particular species of malarial parasite *Plasmodia* are resistant to Chloroquine. So, it explains why Chloroquine is not working in many cases. The criteria, plausibility fits in this case.

### 3.7 Coherence

Available evidence concerning the natural history, biology, and epidemiology of the disease should stick together or form a cohesive whole in the argument for causation.

Example: The rise of smoking in Western countries during early- and mid-twentieth century was accompanied by a corresponding increase in lung cancer mortality [13]. One would expect that the two events are coherent given our current knowledge.

### 3.8 Experimentation

In the event that experimentation such as experimental epidemiological studies, natural experiments, in vitro laboratory experiments, and animal models support of a hypothesis, the observed association is definitely causal. For this reason, some type of epidemiological studies, such as clinical trial, conducted in a controlled

situation, is more supportive of a causal association. In clinical trials, study subjects, who are eligible, are randomly selected to the treatment groups, at least one of which receives the study treatment. The study is often blinded so that patient treatments and follow-up are not biased. The results can prove that the experimental drug can cause the observed clinical outcome or cure of the patient.

### 3.9 *Analogy*

This criterion implies a similarity between things that are otherwise different.

Example: One pharmaceutical drug, such as thalidomide, can cause severe and life-threatening birth defects and death of a baby if the mother or the father is taking the medicine at the time of conception or during pregnancy [14]. Even a single dose of thalidomide can cause serious birth defects in face, heart, eyes, ears, and bones of the baby. Knowing this, one may tend to believe that another drug taken during pregnancy can cause birth defects too. This is called analogy. However, this criterion is more subjective; therefore, this argument is rather weak.

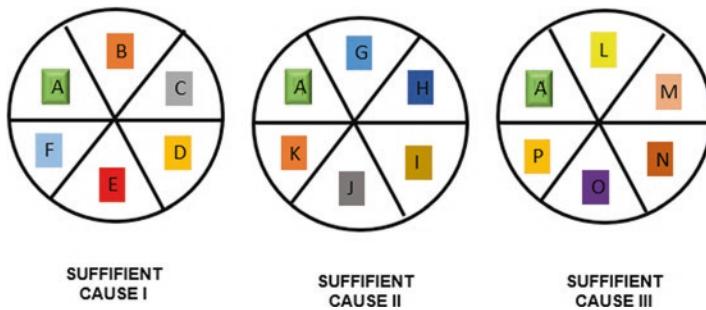
Finally, the proponent of the criteria for causal association, Bradford Hill cautioned that “none of the viewpoints can bring indisputable evidence for or against the cause-and-effect hypothesis” [3]. The Hill’s criteria should be used as a guide for making a judgement for a causal association. More of these criteria work for a situation, more likely the association is causal.

## 4 Rothman’s Causal Pie

In 1976, Kenneth Rothman proposed a conceptual framework for causal theory of diseases. Rothman’s proposed theory of causation is known as the *sufficient-component cause model*—popularly called *Rothman’s causal pie* [15]. As we are experiencing more chronic and multifactorial diseases (such as heart disease, cancer, diabetes, etc.) over time, Rothman’s causal theory better explains diseases of multifactorial origin.

Kenneth Rothman explained the causal factors of a chronic disease as follows:

- Sufficient cause: Several risk factors together make a disease. Combination of all of them is called *sufficient cause*.
- Necessary cause: When one factor is always present to cause a disease, it is called a *necessary cause*. In the example presented in Fig. 10.2, “A” is a necessary cause.
- Component cause: In a multifactorial cause, each of the factors is a component cause.



**Fig. 10.2** A schematic diagram of Rothman's causal pie, modified [13]

#### 4.1 Illustration of Rothman's Pie

Figure 10.2 provides a schematic diagram, where there are three sufficient causes: sufficient cause I, sufficient cause II, and sufficient cause III. In this figure, each sufficient cause contains six factors, indicated by six letters. The letter “A” is a necessary cause, because it is present in all sufficient causes. Each letter of a sufficient cause (such as A, B, C, D, E, and F for the sufficient cause I) is a component cause.

Let us take a practical example—High blood pressure has reached epidemic proportions among the population worldwide. In the USA, one in every three adults is hypertensive and another one in three is pre-hypertensive. Several risk factors contribute to high blood pressure, including:

**BPA:** A chemical called Bisphenol A is used in a wide variety of products such as water bottles, food containers, contact lenses, and even dental fillings. New studies show that BPA exposure can increase blood pressure. The underlying mechanisms may involve alteration of cardiac  $\text{Ca}^{2+}$  handling, ion channel inhibition/activation, oxidative stress, and genome/transcriptome modifications [16].

**Salt:** The amount of salt a person needs in their diet varies based on age and health levels. The United States Department of Agriculture (USDA) recommends that an adult consumes no more than 2.3 g of sodium per day, which is equivalent to 5.75 g (less than one teaspoon) of salt. An increased amount salt intake can disturb the sodium and water balance, causing more fluid retention, an increase in blood volume, and an increase in blood pressure.

**Sugar:** Eating excess of sugar can increase blood pressure. Sugar increases the uric acid levels which block the body’s ability to make nitric oxide. Nitric oxide is necessary to relax your blood vessels, which reduces blood pressure. Without enough nitric oxide, blood vessels will stiffen, causing an increase in blood pressure.

**Sleep apnea:** Most adults need 7–9 hours of sleep. Sleep apnea is a sleeping disorder in which the temporary cessation of breathing (apnea) occurs more often than normal during sleep. It causes lack of oxygen supply to the brain and other organs due to sleep deprivation. Obstructive sleep apnea (OSA) is a more common form of sleep apnea, which is associated with cardiovascular diseases including hypertension [17].

**Thyroid hormones:** With hypothyroidism, body temperature and oxygen consumption are decreased, which can result in increased blood pressure.

**Medications:** Several medications can increase blood pressure. These include nonsteroidal anti-inflammatory drugs (NSAIDs) (Advil, ibuprofen, aleve, Naprosyn, etc.); nasal decongestants (pseudoephedrine and phenylephrine); hormonal birth control pills; and antidepressant drugs (serotonin and dopamine).

**Dehydration:** Chronic dehydration can lead to hypertension.

**Air pollutants:** Recent research suggest that there is a link between high amounts of air pollution with high blood pressure.

**Age:** Blood pressure increases with increasing age.

•

### Box 2

High blood pressure is a combination of many risk factors. Combining all of the risk factors (or at least some of the high-risk factors) make it a sufficient cause. In Fig. 10.1, there are three sufficient causes, meaning that the disease can be caused by a combination of factors within each of the three groups. The factors within each sufficient cause (such as sufficient cause I, or sufficient cause II, or sufficient cause III) are called *component causes*. In the real-life example of high blood pressure, BPA, salt, sugar, sleep apnea, etc. are *component causes*. The combination of component causes make a sufficient cause. In the schematic figure (Fig. 10.1), there is only one component which is common to all the three sufficient causes. It is the letter “A”, which is common to all. It is called a necessary cause, which must be present to cause a disease.

## 5 Further Practice

1. Exposure must come before the disease. What type of Hill’s criteria is this called?
  - (a) Dose-response
  - (b) Temporality
  - (c) Analogy
  - (d) Experimentation
2. Several studies show the same results. What type of Hill’s criteria is this called?
  - (a) Consistency
  - (b) Strength of association
  - (c) Plausibility
  - (d) Coherence
3. If disease A occurs, disease B should occur. What type of Hill’s criteria is this called?
  - (a) Consistency
  - (b) Strength of association
  - (c) Plausibility
  - (d) Coherence

4. The incidence of a disease increases with the increase of exposure. What type of Hill's criteria is this called?

- (a) Dose-response
- (b) Temporality
- (c) Analogy
- (d) Experimentation

5. Pick up the correct answer of the three options and fill up the gap:

- (a) Necessary cause
- (b) Component cause
- (c) Sufficient cause

A. All the factors of a disease combined is called \_\_\_\_\_.

B. The factor must be present to cause a disease \_\_\_\_\_.

C. Many risk factors together make a disease. Each of them separately is called \_\_\_\_\_.

- 6. Rothman's pie is applicable to chronic diseases. True/False
- 7. All the component causes combined is called sufficient cause. True/False
- 8. A cross-sectional study can determine causal relationships. True/False
- 9. A cross-sectional study can identify risk factors. True/False
- 10. A case-control study can determine causal relationships. True/False

### **Short Answers**

11. Define "confounder" with an example.
12. Mention and explain four criteria proposed by Bradford Hill
13. In a cohort study, you can study hazard ratio. Explain with an example what you mean by hazard ratio.
14. Cross-sectional studies cannot establish a causal relationship. Explain.
15. Give an example of a necessary cause, according to Rothman's causal pie.

### **Answers Keys**

1. (b)
2. (a)
3. (c)
4. (a)
- 5A. (c)
- 5B. (b)
- 5C. (b)
6. True
7. True
8. False
9. True
10. True
11. to 15. Review the chapter

## References

- Pardhe BD, Ghimire S, Shakya J, Pathak S, Shakya S, Bhetwal A, Khanal PR, Parajuli NP. Elevated cardiovascular risks among postmenopausal women: a community based case control study from Nepal. *Biochem Res Int.* 2017;2017:3824903. <https://doi.org/10.1155/2017/3824903>.
- El Khoudary SR, Brooke Aggarwal B, Theresa M, Beckie TM, Hodis HN, Johnson AE, Langer RD, et al. Menopause transition and cardiovascular disease risk: implications for timing of early prevention: a scientific statement from the American Heart Association. *Circulation.* 2020;142(25):e506–32. <https://doi.org/10.1161/CIR.0000000000000912>.
- Hill AB. The environment and disease: association or causation? *Proc R Soc Med.* 1965;58:295–300.
- Doll R. Cancer. In: Witts LJ, editor. *Medical surveys and clinical trials: some methods and applications of group research in medicine.* 2nd ed. London: Oxford University Press; 1964. p. 333.
- Leyland KM, Judge A, Javaid MK, Diez-Perez A, Carr A, Cooper C, et al. Obesity and the relative risk of knee replacement surgery in patients with knee osteoarthritis: a prospective cohort study. *Arthritis and Rheumatology.* 2016;68(4):817–25. <https://doi.org/10.1002/art.39486>.
- Barreto PR, Diniz DLO, Lopes JP, Barroso MC, Daniele TMDC, et al. Obstructive sleep apnea and wake-up stroke—a 12 months prospective longitudinal study. *J Stroke Cerebrovasc Dis.* 2020;29:104564.
- Mitra AK, Engleberg NC, Glass RI, Chowdhury MK. Fatal dysentery in rural Bangladesh. *J Diarrheal Dis Res.* 1990;8:12–7.
- Choudhury A, Routray D, Swain S, Das AK. Prevalence and risk factors of people at-risk of obstructive sleep apnea in a rural community of Odisha, India: a community based cross-sectional study. *Sleep Med.* 2019;58:42–7. Available online: <https://pubmed.ncbi.nlm.nih.gov/31078079/>
- Mitra AK, Mahalanabis D, Ashraf H, Unicomb L, Eeckels R, Tzipori S. Hyperimmune cow colostrum reduces diarrhoea due to rotavirus: a double-blind, controlled clinical trial. *Acta Paediatr.* 1995;84:996–1001.
- Sarker SA, Casswall TH, Mahalanabis D, Alam NH, Albert MJ, Brussow H, et al. Successful treatment of rotavirus diarrhea in children with immunoglobulin from immunized bovine colostrum. *Pediatr Infect Dis J.* 1998;17(12):1149–54.
- Mitra AK, Bose BK, Kabir H, Das BK, Hussain M. Arsenic-related health problems among hospital patients in southern Bangladesh. *J Health Popul Nutr.* 2002;20:198–204.
- Mitra AK, Faruque FS. Breast cancer incidence and exposure to environmental chemicals in 82 counties in Mississippi. *South Med J.* 2004;97(3):259–63.
- Islami F, Torre LA, Jemal A. Global trends of lung cancer mortality and smoking prevalence. *Transl Lung Cancer Res.* 2015;4(4):327–38. <https://doi.org/10.3978/j.issn.2218-6751.2015.08.04>.
- Vargesson N. Thalidomide-induced teratogenesis: history and mechanisms. *Birth Defects Res C Embryo Today.* 2015;105(2):140–56. <https://doi.org/10.1002/bdrc.21096>.
- Rothman KJ, Greenland S. Causation and causal inference in epidemiology. *Am J Public Health.* 2005;95:S144. <https://doi.org/10.2105/AJPH.2004.059204>.
- Gao X, Wang HS. Impact of bisphenol a on the cardiovascular system - epidemiological and experimental evidence and molecular mechanisms. *Int J Environ Res Public Health.* 2014;11(8):8399–413. <https://doi.org/10.3390/ijerph110808399>.
- Mitra AK, Bhuiyan AR, Jones EA. Association and risk factors for obstructive sleep apnea and cardiovascular diseases: a systematic review. *Diseases.* 2021;9:88. <https://doi.org/10.3390/diseases9040088>.

# Chapter 11

## Bias, Confounding, and Effect Modifier



Dipak Kumar Mitra and Abdullah H. Baqui

### Learning Objectives

After completing this chapter, you will be able to:

- Understand the basic concepts of bias
- Identify different types of information bias
- Differentiate between selection bias and information bias
- Identify different types of selection bias
- Use methods of minimizing bias and confounding

## 1 Introduction

Epidemiological procedures require the measurement of variables (outcome and exposure variables) to reach a conclusion. Any measurement procedure that is subject to an error may produce an incorrect estimate of absolute measures (prevalence and incidence) and measures of association between exposure and outcome (risk ratio, odds ratio, risk difference, etc.). An error may be due to chance (random error) or due to a systematic flaw in the measurement procedure or study design. Systematic error is commonly known as bias. In this chapter, we will discuss bias in measures of association between exposure and outcome variables.

---

D. K. Mitra (✉)

Department of Public Health, School of Health and Life Sciences, North South University,  
Dhaka, Bangladesh

e-mail: [dipak.mitra@northsouth.edu](mailto:dipak.mitra@northsouth.edu)

A. H. Baqui

Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA  
e-mail: [abaqui@jhu.edu](mailto:abaqui@jhu.edu)

Definition of bias: Bias can be defined as a *systematic error* in measuring the association between an outcome of interest and an exposure of interest using data collected from a sample. For example, if the true association between lung cancer and smoking results in an odds ratio (OR) of 6.5 in a population without any bias, and if a research team observes this association with an OR of 9.5 using data from a sample of this population, the observed estimate could be an overestimate due to bias. The amount of bias is the difference between the true association and the observed association. The direction of bias can be either away from the null value (overestimates the true association) or toward the null (underestimates the true association). Any measurement without any bias is called a valid estimate. In other words, lack of validity is bias and lack of bias is valid.

## 2 Sources of Bias

Bias in estimating an association between exposure of interest and outcome of interest can result from three different ways.

### 2.1 Selection Bias

If we can collect data from the entire target population, there is no selection of participants and no chance of selection bias. However, when we estimate the association between exposure and outcome using data collected from a randomly selected sample of the target population, bias may or may not happen. Ideally, the distribution of outcome and exposures in a random sample is expected to be similar to that in the population. If this assumption is true, the observed measure of the association will be no different from the true measure in the population, meaning “*no biases*.” However, the selection of samples may be employed in a way that the distribution of exposure and outcome variables in the sample is not similar to that in the population. In that case, the observed measures of association will be biased. Selection bias is important for the generalizability of the study results, which means the results with selection bias are not generalizable to the target population from which the sample was drawn [1]. This is also known as lack of “*external validity*.” The magnitude and direction of selection bias are unpredictable.

#### 2.1.1 Reasons for Selection Bias

1. Selection bias can occur when nonprobability sampling procedures (voluntary and convenience sampling) are used in the study.
2. Inclusion and exclusion criteria are not strictly defined and followed.
3. High refusal of the participants to participate in the study occurs.

4. High lost to follow-up of the participants in cohort and experimental studies.
5. Bias due to the development of competing risk before the event of interest in a cohort study.
6. Berkson's bias: This may not be a problem in a population-based case-control study. In a hospital-based case-control study, both the cases and the controls should be selected from the same study-base from where the cases developed. When both cases and controls are selected from a hospital, selection bias can occur for both case and control selection; the problem in control selection is likely to be higher. This bias is also known as Berkson's bias. In 1946, Joseph Berkson described this bias in the assessment of the association between an exposure and a disease due to the conduct of the study in a hospital, where hospital admission was affected by both exposure and disease. Exposed persons are more likely to be hospitalized compared to nonexposed persons. On the other hand, the probability of hospital admission between case disease and control disease may also differ, resulting in bias. For example, let's investigate an association between coronary heart disease (CHD) and a risk factor of hypertension. We select patients admitted with CHD as cases and people admitted with another disease as controls. The admission probability of the two diseases may be different as hypertension (risk factor) has an independent probability of admission. Thus, our study will likely get more hypertensive cases resulting in Berkson's bias. Berkson's bias overestimates the true association.
7. Diagnostic bias: Symptomatic individuals with known risk factors are more likely to undergo diagnostic procedures compared to individuals with symptoms but without the risk factors, resulting in identification of more exposed cases than unexposed cases. For example, smokers with symptoms of coughing blood may more often go for diagnostic procedures for lung cancer compared to those with the symptoms of coughing blood, but not smokers. In a case-control study, this will overestimate the association between smoking and lung cancer (odds ratio).
8. Survival bias: Cases who do not survive to be included in the study are likely to be exposed to more severe lethal risk factors. Thus, fewer exposed cases are selected in the study, resulting in an underestimate of the true association between the exposure and disease.

### 2.1.2 Minimizing Selection Bias

Selection bias can be minimized by using appropriate selection procedures of the samples from the target population.

1. Random selection of participants: The best approach to minimize selection bias is simple random sampling of participants from the target/source population, where each of the participants in the population has a non-zero, equal, and independent probability of selection. This will ensure external and internal validity, provided no other biases (confounding and information bias) exist.

2. Randomized allocation of exposure/intervention in experimental studies eliminates the selection bias for internal validity. External validity may not be fully ensured by randomization in randomized controlled trials (RCTs) due to many other factors, such as settings of participant recruitment (community and primary/secondary/tertiary level health facility), refusal to participate, noncompliance, and loss to follow-up.
3. Use of strict inclusion and exclusion criteria for the selection of study participants, especially in a case-control study.
4. Minimize refusal and loss to follow-up with the participants.

### 3 Information Bias

Information bias occurs due to measurement error in assessing exposures and outcome variables in the study participants. Based on the measurement of exposure and outcome, we classify participants into two groups: (1) diseased (with an outcome) and not diseased (without outcome) and (2) exposed and unexposed. The consequence of measurement error is the misclassification of exposure and outcome status, and eventually, it produces biased measures of association. The extent of bias will depend on the nature and extent of misclassification in measuring the exposure and the outcome. The nature of misclassification can be false-negative—a true diseased person is classified as not diseased or a true exposed person is classified as unexposed—or false-positive—a true not diseased person is classified as diseased or a true unexposed person is classified as exposed. The extent of misclassification is the proportion of misclassification in either way. The extent of bias due to misclassification may be further compounded by the uniformity of error in two groups of participants, that is, outcome misclassification by exposure category and exposure misclassification by outcome category. If the nature and extent of misclassification are comparable in the two groups, it is known as nondifferential misclassification. It is known as differential misclassification if they are different in the two groups. Nondifferential misclassification can be ensured using the same measurement tools and personnel for both groups. If we can assume a nondifferential misclassification, the bias will always be toward the null (or underestimation). For differential misclassification, the direction of bias is unpredictable. The magnitude of bias in both situations will depend on the extent of misclassification. Information bias in epidemiological studies occurs due to (1) errors in respondents' responses, (2) use of faulty measurement instruments, and (3) errors caused by observers (or data collectors) [2].

### 3.1 Examples of Information Bias in Epidemiological Studies

#### 3.1.1 Recall Bias

Recall bias is commonly encountered in case–control studies, where reported exposure history from cases and controls is collected. Although both cases and controls are likely to have recall issues, cases are usually more likely to have more accurate recall compared to controls. This results in nondifferential misclassification of exposures between cases and controls. For example, suppose the research goal is to identify the association between smoking statuses in lung cancer using a case–control study. In that case, the cases with lung cancer are more likely to remember and report of their smoking statuses more accurately than those without lung cancer (controls), resulting in a recall bias.

#### 3.1.2 Minimizing Information Bias

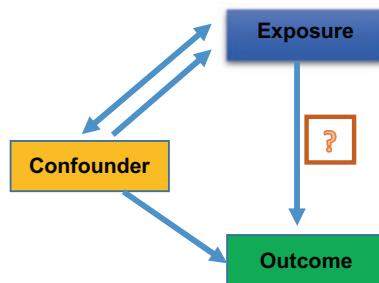
1. Use of standardized measurement instruments
2. Administer instruments equally to cases and controls to ascertain exposure
3. Use multiple sources of information
  - Questionnaires
  - Direct measurements
  - Patient registries
  - Case records
4. Use multiple sets of controls

## 4 Confounding Bias

The term “*confounding*” refers to a situation in which a noncausal association between the exposure of interest and the outcome of interest is observed because of the influence of a third factor (or a variable), usually designated as a confounding variable or simply a confounder (Fig. 11.1). Confounding is a population characteristic and should not be evaluated from the sample. The variable which is a confounder must meet the following criteria:

- **Must be an independent cause or risk factor of the outcome of interest.** This means that the association between the confounding variable (e.g., drinking alcohol) and the outcome (e.g., bladder cancer) is present in both exposed (e.g., smokers) and unexposed populations (e.g., nonsmokers).
- **Must be causally or noncausally associated with the exposure of interest.** That means the confounding variable (e.g., drinking alcohol) must be more common or less common in the exposed population (e.g., smokers) than the unex-

**Fig. 11.1** Effect of confounding



posed population (e.g., nonsmokers). When the confounding variable is associated with the exposure, the distribution of the confounding variable becomes imbalanced between the exposed and unexposed population resulting in confounding.

- **Cannot be an intermediate variable in the causal pathway between exposure and the outcome.** This means that the confounding variable (e.g., drinking alcohol) cannot be an intermediate step or a moderator in the causal pathway between the exposure (e.g., smoking) and the outcome (e.g., bladder cancer).

Confounding is a population characteristic that is presented in Fig. 11.1.

#### 4.1 Assessment of Confounding

The presence of confounding should be assessed based on the *prior* knowledge about the common association of the potential confounding variable with the exposure and the outcome. In addition to the *prior* knowledge, the presence of confounding can be assessed in the study data by examining whether the potential confounding variable is associated with both exposure and outcome.

As confounding produces a biased association between exposure and the outcome, it must be removed to obtain the true magnitude of the causal association between the exposure and the outcome. The following approaches are used to eliminate confounding in an epidemiological study.

- A. In the design stage
  - (a) Randomization
  - (b) Restriction
  - (c) Matching
- B. In the analysis stage
  - (a) Stratified analysis
  - (b) Multiple regression

#### 4.1.1 Randomization

Randomized allocation can eliminate confounding for all known, unknown, measured, or unmeasured confounding variables; thus, this is the best approach to addressing confounding, if possible. As in the randomized allocation or randomized controlled trial (RCT), the exposure of interest is allocated randomly and no association between the exposure and the potential confounding variables can exist in RCTs [3].

Randomized allocation eliminates confounding (Fig. 11.2).

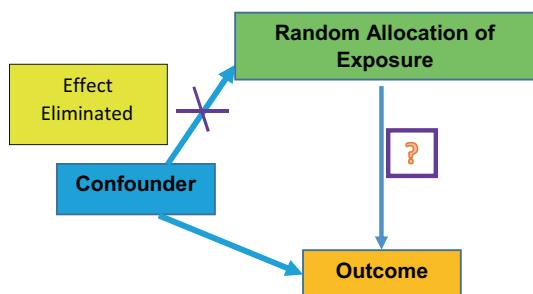
#### 4.1.2 Restriction

Restriction is the approach to restrict the study population to only one category of participants with respect to the confounding variable. For example, if alcohol drinking is a potentially confounding variable, the study is conducted only either for drinkers or nondrinkers. This approach prevents the imbalance of the confounding variable between the exposed and unexposed population, as the study includes only one group of participants. Although this eliminates the confounding, it is rarely done as the study results are only applicable to people included in the study [4].

#### 4.1.3 Matching

Matching is commonly done in case-control studies to control a few important known confounders. Confounding variable status (presence or absence) is matched between cases and controls. Matching variables that are not confounding may not be useful; thus, matching should be done carefully. Although matching eliminates the confounding for matched variables, it introduces a type of selection bias (as controls are no longer selected randomly), which should be accounted for in the analysis using stratified or conditional logistic regression analysis.

**Fig. 11.2** Effect of confounding eliminated



#### 4.1.4 Stratified Analysis

Stratified analysis can eliminate confounding from all measured variables in a particular study. Confounding due to unmeasured variables cannot be addressed in this approach. In the stratified analysis, the following analytic steps are followed.

- (a) Measure of association between the exposure and the outcome is estimated in the first step ignoring the confounding status of the participants. This is known as crude or unadjusted analysis. The measure of association, for example, odds ratio, is thus known as crude/unadjusted odds ratio.
- (b) In the next step, the study participants are divided into the categories of confounding variables, ignoring the exposure and outcome status. Each group is known as a stratum. Then, measures of association between exposure and outcome are estimated separately for each stratum. If the stratum-specific estimates seem different, no further analysis is done and separate stratum-specific estimates are reported. This is considered as *effect modification*. But if the stratum-specific estimates look similar, then the estimates are *pooled* using the *Mantel-Haenszel weights* to obtain a pooled or combined estimate. This is also known as an *adjusted estimate*. Then, we compare the crude and adjusted estimate. If both look very similar, we conclude that there is no confounding and in that case, we report the crude estimate. However, if the adjusted estimate is >10% different from the crude estimate, we conclude that there is a confounding effect and report the adjusted estimate. Usually, researchers report both crude and adjusted estimates for the readers.

It is important to note that in the stratified analysis, in each stratum, the confounding variable is constant; thus, it cannot confound the exposure/outcome association as the confounding variable cannot vary across the exposure status.

#### 4.1.5 Multiple Regression Analysis

Multiple regression (linear, logistic regression, etc.) is often employed to adjust for confounding variables in observational studies. Multiple regression analysis assumes that all potentially confounding variables are constant between the exposed and unexposed people, thus eliminating the confounding effect. This method is similar to stratified analysis, where each stratum's confounding variable is constant.

## 5 Types of Confounding Effect (Positive, Negative, and Qualitative)

Confounding effect is positive when it overestimates the true magnitude of association; it is negative when it underestimates the true magnitude of association; and the confounding effect is qualitative when the confounding reverses the direction of the association. A hypothetical example is shown in Table 11.1.

## 6 Effect Modification

Effect modification, also known as *interaction*, is a situation when two or more risk factors modify the effect of each other on the outcome of interest. An effect modifier variable is also considered a third variable concerning the association between the exposure and the outcome of interest, but it is different from the confounding. Confounding results from the interrelation of multiple risk factors with the outcome, which is not a real causal effect, and we want to eliminate the confounding effect from our results. On the other hand, effect modification is a real biological effect and we want to identify or establish the effect modification and report it to the readers. If an effect modification is present, the effect of the exposure on the outcome in the presence of the effect modifier differs from that in the absence of the effect modifier [5]. For example, if smoking is an effect modifier of the effect of asbestos on lung cancer, the association between asbestos and lung cancer among smokers will be different from that among nonsmokers. One important issue for the effect modification is the measurement scale, such as an additive scale or a multiplicative or a ratio scale. Although the assessment of effect modification is the same for both scales, the results may differ between the scales in the same study. In the same data, effect modification may be present in the additive scale but not in the ratio scale and vice versa, as illustrated using data in the next few tables (Tables 11.2a, 11.2b, and 11.2c).

Further, effect modification can be classified into synergistic, antagonistic, quantitative, and qualitative. If the presence of the effect modifier enhances the effect of

**Table 11.1** Different types of confounding effect: a hypothetical example

Example	Crude estimate of OR	Adjusted estimate of OR	Type of confounding
1	3.5	1.0	Positive
2	4.6	2.3	Positive
3	1.0	2.2	Negative
4	0.3	0.7	Positive
5	0.8	0.3	Negative
6	2.5	0.6	Qualitative
7	0.6	2.4	Qualitative

**Table 11.2a** Effect modification additive and ratio scales

	Effect modifier		Remarks
Exposure status	Present	Absent	
Risk in exposed	0.04	0.01	
Risk in not exposed	0.004	0.001	
Risk difference	0.036	0.009	Effect modification present
Risk ratio	10	10	No effect modification

**Table 11.2b** Effect modification additive and ratio scales

	Effect modifier		Remarks
Exposure status	Present	Absent	
Risk in exposed	0.30	0.10	
Risk in not exposed	0.05	0.05	
Risk difference	0.25	0.05	Effect modification present
Risk ratio	6	2	Effect modification present

the exposure of interest, then the modifier and the exposure are called synergistic (positive interaction). If the effect modifier diminishes the effect of the exposure, then the modifier and the exposure are called antagonistic (negative interaction).

If the measures of association (e.g., OR) between exposure and outcome are in the same direction (either both are positive or both are negative), it is known as quantitative effect modification (Table 11.3); and if they are in a different direction (one positive and the other negative or vice versa), it is known as qualitative effect modification (Table 11.4).

Both confounding and effect modification are assessed using stratified analysis. This can be presented in the following flow chart (Fig. 11.3).

Figure 11.3 shows the identification of confounding/effect modification step by step. We start with the crude analysis ignoring the confounding or effect modification by the third variable and estimate the crude measure of association. In the second step, we conduct a stratified analysis and assess whether the stratum-specific estimates are similar. If they are not similar, we stop the analysis and conclude that effect modification is present. We report stratum-specific associations. However, if the stratum-specific estimates are similar, we pool the estimates using Mantel-Haenszel weights and obtain the combined/adjusted measure of association. We then compare the adjusted estimate with a crude estimate. If the adjusted estimate differs by more than 10%, we conclude that there is evidence of confounding and report the adjusted estimate. If the difference between the crude and adjusted estimates is less than 10%, we conclude that there is no confounding and report the crude estimate. In common practice, researchers report crude and adjusted estimates for the readers.

**Table 11.2c** Effect modification additive and ratio scales

	Effect modifier		Remarks
Exposure status	Present	Absent	
Risk in exposed	0.20	0.50	
Risk in not exposed	0.10	0.40	
Risk difference	0.10	0.10	No effect modification
Risk ratio	2	1.25	Effect modification present

**Table 11.3** Quantitative effect modification

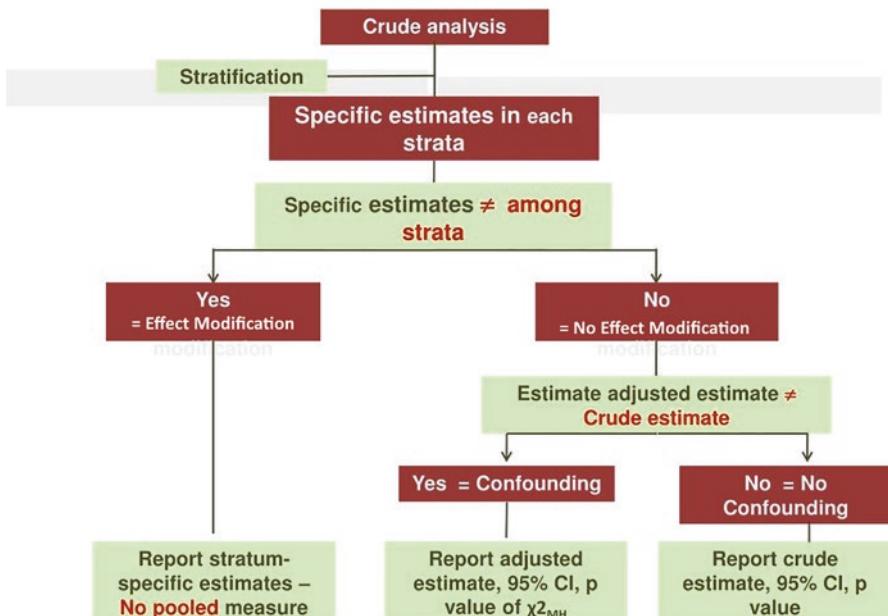
Family history of clubfoot	Maternal smoking	Cases	Controls	Stratified OR for maternal smoking
Yes	Yes	14	7	3.64
	No	11	20	
No	Yes	118	859	1.45
	No	203	2143	

## 7 Further Practice

1. There is no chance of selection bias if we can collect data from the entire target population.      True/False
2. Bias is a random error.      True/False
3. The results with selection bias are not generalizable to the target population from which the sample was drawn. This is also known as lack of:
  - (a) External validity
  - (b) Internal validity
  - (c) Both internal and external validity
4. The method of selecting samples by convenience sampling can cause
  - (a) Selection bias
  - (b) Information bias
  - (c) Both selection bias and information bias
5. Refusal of the participants to participate in the study can result in:
  - (a) Selection bias
  - (b) Information bias
  - (c) Both selection bias and information bias
6. Berkson's bias is a type of:
  - (a) Selection bias
  - (b) Information bias
  - (c) Both selection bias and information bias

**Table 11.4** Qualitative effect modification

Smoking	Caffeine	No. pregnancies	Delayed conception >12 months	Stratified OR <sub>A</sub>	95% CI
No	No	575	47	2.62	1.36–4.98
	≥ 301 mg/d	90	17		
Yes	No	76	15	0.62	0.27–1.45
	≥301 mg/d	83	11		

**Fig. 11.3** Flow diagram to assess confounding and effect modification

7. Information bias occurs due to measurement errors in the assessment of exposures and outcome variables.      True/False
8. Recall bias is commonly encountered in:
  - (a) Case-control study
  - (b) Cross-sectional study
  - (c) Cohort study
9. You can reduce information bias if:
  - (a) The number of cases and controls are the same
  - (b) The number of cases is greater than that of controls
  - (c) The number of controls is multiple (two to three per case)

10. To get rid of confounding in the study design phase, use the following, except:
- (a) Randomization
  - (b) Matching
  - (c) Restriction
  - (d) Plan for stratified analysis
11. If alcohol drinking is a potential confounding variable, the study should be conducted only either in drinkers or nondrinkers to avoid the confounding effect.      True/False
12. If smoking is an effect modifier of the effect of asbestos on lung cancer, the association between asbestos and lung cancer among smokers and among non-smokers will be:
- (a) Similar
  - (b) Different
  - (c) No change
13. If the measures of association (e.g., OR) between exposure and outcome are in the same direction (either both are positive or both are negative), it is known as:
- (a) Quantitative effect modification
  - (b) Qualitative effect modification
  - (c) Positive effect modification
  - (d) Negative effect modification
14. RCT can eliminate confounding      True/False
15. Matching eliminates the confounding for matched variables, however, it introduces a type of selection bias, why?

### Answer Keys

1. True
2. False
3. (a)
4. (a)
5. (a)
6. (a)
7. True
8. (a)
9. (c)
10. (d)
11. True
12. (b)
13. (a)
14. True
15. Because controls are no longer selected randomly.

## References

1. Simundić AM. Bias in research. *Biochem Med (Zagreb)*. 2013;23(1):12–5. <https://doi.org/10.11613/bm.2013.003>.
2. Pannucci CJ, Wilkins EG. Identifying and avoiding bias in research. *Plast Reconstr Surg*. 2010;126(2):619–25. <https://doi.org/10.1097/PRS.0b013e3181de24bc>.
3. Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench*. 2012;5(2):79–83.
4. Jager KJ, Zoccali C, MacLeod A, Dekker FW. Confounding: what it is and how to deal with it. *Kidney Int*. 2008;73(3):256–60.
5. Corraini P, Olsen M, Pedersen L, Dekkers OM, Vandenbroucke JP. Effect modification, interaction and mediation: an overview of theoretical insights for clinical investigators. *Clin Epidemiol*. 2017;9:331–8. <https://doi.org/10.2147/CLEPS129728>.

# Chapter 12

## Epidemic Investigation and Control



Rajat Das Gupta and Sanjoy Kumar Sadhukhan

### Learning Objectives

After completing this chapter, you will be able to:

- Describe the differences between epidemic, outbreak, pandemic, and endemic
- Describe different steps of outbreak investigation
- Understand the future preventive measures
- Understand the context by comparing with a real-life example

## 1 Overview

Epidemic investigation is a major task in the infectious disease epidemiology. In order to properly investigate and control an epidemic, systematic approach is required. The fundamental principles of epidemic investigation and control is discussed in this chapter. The concept of epidemic, outbreak, endemic, pandemic, cluster, epizootic, attack rate, and case fatality rate is discussed. A step-by-step method of epidemic investigation is discussed along with statistical approaches. Finally, the chapter provides a real-life example of the concept of a makeshift hospital by involving the community people in the control, a large diarrheal epidemic in rural Bangladesh.

---

R. D. Gupta (✉)

Department of Epidemiology and Biostatistics, Arnold School of Public Health, University of South Carolina, Columbia, SC, USA  
e-mail: [rajatdas@email.sc.edu](mailto:rajatdas@email.sc.edu)

S. K. Sadhukhan

Department of Epidemiology, All India Institute of Hygiene & Public Health, Bidhan Nagar Campus, Bidhan Nagar (Salt Lake), Kolkata, India

## 2 Useful Terms

### 2.1 *Epidemic*

According to the Centers for Disease Control and Prevention (CDC), epidemic refers to a situation where there is an increase in the number of cases of a disease in a population, which is above than the normally expected number of cases of the disease among the population of the same area at a given time [1]. For example, smallpox has been declared eradicated in 1980. If one single case of smallpox is detected around the world, it is clearly above than the normally expected number of cases and will be declared as an epidemic.

### 2.2 *Outbreak*

Outbreak has the same definition of epidemic, but it is used in a smaller scale or in a limited geographical area compared to an epidemic which is used in a wider scale [1].

### 2.3 *Pandemic*

When an epidemic spreads over to several countries or continents, it is called a pandemic [1]. The global pandemic of coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has been spread across the continents and is a classic example of a pandemic. Similarly, human immunodeficiency virus (HIV) infection causing acquired immune deficiency syndrome or AIDS is currently a worldwide pandemic disease.

### 2.4 *Endemic*

If a disease is constantly/habitually present in low rates in a community or in a geographical area, then it is called endemic [1]. For example, tuberculosis is an endemic disease in Bangladesh; flu is endemic in many countries.

## 2.5 *Hyperendemic*

The term hyperendemic disease is used to refer a disease which is constantly and persistently present in a population as a high rate in a given time. For example, at the present time, dengue fever is hyperendemic in Bangladesh and in India.

## 2.6 *Cluster*

When the cases of an event/disease aggregates in space and time, then it is called as a cluster, e.g., the number of cases of childhood leukemia near to a radiation site.

## 2.7 *Epizootic*

An epidemic disease affecting nonhumans is called epizootic. But it will be called an epidemic if the same disease affects human at a large scale. For example, the outbreak of H5N1 (a highly infective strain of avian influence virus) in bird population is called epizootic. In case of an outbreak of H5N1 bird flu in humans in a larger scale, it is called an epidemic.

## 2.8 *Attack Rate*

Attack rate can be defined as the number of new cases of disease during the specified time period in a specific area divided by the number of population at risk in that specific area during the same time. This is also known as incidence proportion (applicable for short duration illness.)

## 2.9 *Secondary Attack Rate*

Secondary attack rate can be defined as the number of new cases among the susceptible contacts of the primary case divided by the total number of contacts [2].

## 2.10 Case Fatality Rate (CFR)

Case fatality rate may be defined as the proportion of deaths from a specific disease among all the individuals who are diagnosed with that disease during that specified time period in that specific area [3].

## 3 Threshold Level of Outbreak (for Epidemic Prone Diseases)

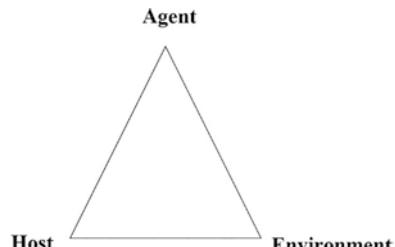
It varies from disease to disease and place to place. For some diseases or situations, pre-existing thresholds at national or international levels exist, e.g., a single case of smallpox anywhere in the world, acute flaccid paralysis in an area where polio has been eliminated, measles in a tribal area, etc. For most of the common communicable diseases, it is based on the past or historical data. For example, monthly mean incidence of past three years for diarrhea, typhoid, malaria, etc. In other places, in the absence of such information, simply the increasing number of cases or deaths over a short period (weeks) could be considered as a trigger factor for diseases such as diarrhea, hepatitis, malaria, dengue, avian influenza, etc.

The basic issue in an outbreak is an imbalance of the “Epidemiological Triad” of agent, host, and environment (Fig. 12.1). The imbalance occurs when there is some change in the way these three components interact. These changes need to be understood well before appropriate actions can be undertaken [4].

## 4 Outbreak Trigger Levels for Responses

Depending on the magnitude and spread of an outbreak, appropriate responses need to be implemented at different levels for its control. Every country has got its own norms for such responses. The responses suggested to be undertaken at different levels as per trigger levels in a developing country, India, are provided as an example (Table 12.1):

**Fig. 12.1** Epidemiological triad



**Table 12.1** Trigger levels, their significance, and level of response of outbreak used in India [5]

Trigger	Significance	Level of response
I	Suspected or limited outbreak	Local response, e.g., to be managed locally by health worker(s) and medical officers.
II	Outbreak	Both local and district responses, e.g., to be managed by district surveillance officer and rapid response team in addition to local officials.
III	Confirmed outbreak	Local, district, and state-level responses required.
IV	Widespread epidemic	State-level response in addition to local and district levels.
V	Disaster response	Responses at all levels including central level and by partners (international).

In addition to trigger level III and above, state-level responses are specifically asked for when there is an unusual outbreak of unknown etiology especially when the case fatality ratio is high.

#### **4.1 Rapid Response Team (RRT)**

Starting from the district level, rapid response teams function for the control and prevention of outbreak and epidemics. The usual members of such team include one epidemiologist/public health specialist, leading the team; one clinician (specialty varies as per situation, and the nature of outbreak); and one microbiologist. When there is an outbreak of a vector-borne disease (such as malaria or dengue fever), one entomologist is also an important member of the team. Depending on the situation, the team may also include one laboratory technician for sample collection and one health worker to assist the team in data collection. The members of an RRT are healthcare officials and nurses who are assembled on ad hoc basis to compose the team.

The primary role of RRT is to confirm and investigate an outbreak regarding its causative agent, source and/or reservoir, and the mode of transmission. In addition, the team assists the local health system in controlling the outbreak and suggests preventive measures for the future outbreak(s) in the locality. The primary responsibility of controlling and preventing outbreak(s) remains with the respective local health department or the district-level health system.

## 5 Steps of an Outbreak Investigation

According to CDC, the following steps should be taken during an outbreak investigation [6]. Although the chronology of these steps need not to be maintained in all cases, several steps can be taken simultaneously or may be in a different order based on the situation.

### Box 12.1: Steps of an Outbreak Investigation

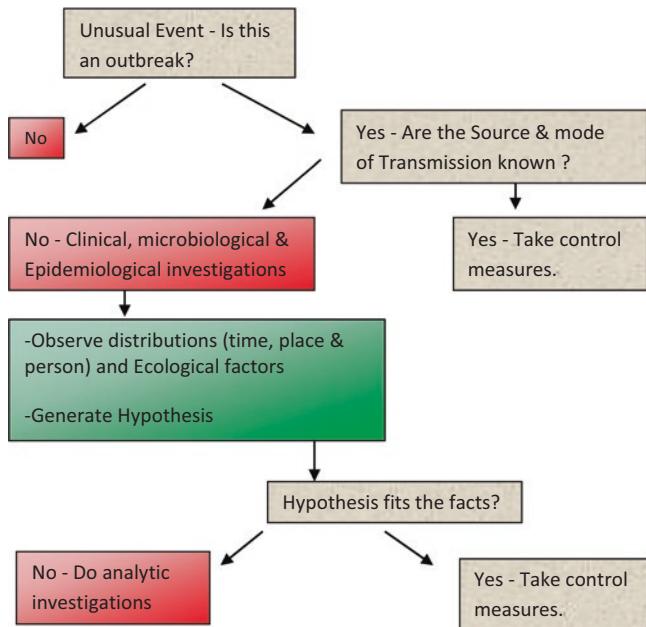
1. Prepare for field work
2. Establish the existence of an outbreak
3. Verify the diagnosis
4. Construct a working case definition
5. Find cases systematically and record information
6. Perform descriptive epidemiology
7. Develop hypotheses
8. Evaluate hypotheses epidemiologically
9. As necessary, reconsider, refine, and re-evaluate hypotheses
10. Compare and reconcile with laboratory and/or environmental studies
11. Implement control and prevention measures
12. Initiate or maintain surveillance
13. Communicate findings
14. Recommend future control measures

Source: Principles of Epidemiology, Lesson 6, Section 2 (Centers for Disease Control and Prevention) [6].

A schematic diagram (Fig. 12.2) shows different steps of an outbreak investigation.

### 5.1 Prepare for Field Work

Field work is an important step in outbreak investigation. The preparation of field work can be divided into two categories: (a) scientific and investigative issues and (b) management and operational issues. A smooth outbreak investigation warrants good preparation in both categories. The whole team should be trained to perform all the steps. They discuss on a daily basis about the problems encountered and in taking a quick decision if plans have to be changed. They should also involve the community people in performing the investigation and in combating the epidemic [6].



**Fig. 12.2** A schematic diagram of different steps of an outbreak investigation

## 5.2 Establish the Existence of an Outbreak

As mentioned previously, the difference between epidemic and outbreak is the extent of the disease. Epidemic is more like a crisis situation, which needs to be investigated and controlled as quickly as possible. It is an important task for the outbreak investigators to verify the true existence of an epidemic. This can be done by comparing the number of cases in an area at the time of investigation with the number of cases from the previous few weeks/months/years. For example, in order to establish a cholera outbreak in a region in a particular time, we need to compare the number of reported cases during that time with the number of cases of the previous years at the same time. In the case of a locality where there is poor documentation of records, alternative methods are taken to verify the number of cases including an interview with the local physicians/public health officials or doing a quick community survey.

Verification is needed to establish whether the increase in cases was attributable to the true outbreak or due to other reasons. Other reasons for the outbreak investigation include strengthening the surveillance system for the disease, introduction of new diagnostic tools, changes in the case definition, and changes in the demographic characteristics of that particular area, including the population size [6].

### ***5.3 Verify the Diagnosis***

This step of verifying the diagnosis is closely related to the establishment of the existence of an outbreak. This step is very important to confirm the proper identification of the disease and in ruling out of any laboratory errors for number of cases.

In order to verify the diagnosis, the investigators first need to review the clinical findings and the laboratory results. Any type of inconsistencies should be verified by a qualified laboratory scientist. Adequate number of specimens should be collected to reconfirm the laboratory results. Second, clinical history can be taken from the patients if possible. This is helpful in generating hypothesis regarding the disease etiology and its spread. Third, frequency distributions can be used to summarize the findings. This helps to characterize the disease spectrum, diagnosis verification, and constructing the case definition [6].

### ***5.4 Construct a Working Case Definition***

A working case definition should be developed for the outbreak investigation. It is restricted by time (e.g., individuals within 3 months of the start of the illness), place (e.g., within five administrative districts of a country), and person (e.g., persons with no previous history of a positive serological test). The case definition should be objective and measurable (e.g., body temperature  $\geq 38$  degree Fahrenheit). The epidemiologists try to reduce the number of false-positive cases.

For “case definition,” it is always better to use a standard definition which should be available from the national public health authority. For diseases of unknown etiology, a suitable “operational definition” is to be made considering the already known clinical features of similar disease in other areas at the present time. Most of the countries in the world maintain standard case definitions for their epidemic-prone diseases. For example, the World Health Organization defines “watery diarrhea” as “three or more watery stools in 24 hours” [6].

### ***5.5 Find Cases Systematically and Record Information***

The reported cases can be a small part of the total number of the cases. For example, for each case of cholera reported in a hospital, 10 more cases are expected to remain in the community. That is why outbreak investigation team needs to find out the total extend of geographical areas affected by the outbreak and find out the total number of cases in those areas. The investigators can conduct a surveillance of cases for case-finding [6].

**Box 12.2: Types of Surveillance**

In epidemiology, surveillance is an ongoing and systematic collection, analysis, and interpretation of health data in the process of describing and monitoring a health event. Surveillance can be conducted either by an active process or a stimulated or enhanced passive process for case detection. In an active surveillance, the investigators actively search for the cases by in-person investigation of households in a community/facility visit or by using a telephone survey. In a stimulated or enhanced passive surveillance, the investigators send letters, describing the scenarios and ask for reports of similar cases in the community. Sometimes the scenario of outbreak is declared to the community people through mass media, and the investigators ask for public assistance in identifying the cases [6].

Contact tracing is done for several reasons: (i) prevent potential further spread of the infection by following up the exposed and infected contacts; (ii) provide case management for ongoing patients and prevent further spread from the primary cases to other (which is called secondary attack rate); and (iii) find out the source of infection [6, 7].

## **5.6 Perform Descriptive Epidemiology**

The epidemiological characteristics of the person who were affected by the epidemic should be described in time, place, and person.

### **5.6.1 Time**

To describe the epidemic in terms of time, an epidemic curve is used. Epidemic curve is a histogram which depicts the course of the epidemic with time. In the x-axis, the date and time of the onset of epidemic are shown. In the y-axis, the number of cases (and deaths) is shown [8].

### **5.6.2 Place**

A spot map can be used to show the residence, working address, or place of exposure of the cases. A cluster pattern indicates a common place of exposure (i.e., a restaurant). An area map with area-specific rates are useful to compare the case rates across the areas [8]. Environmental samples are also collected and analyzed in a laboratory to find out any environmental source of the epidemic, such as a contaminated water source.

### 5.6.3 Person

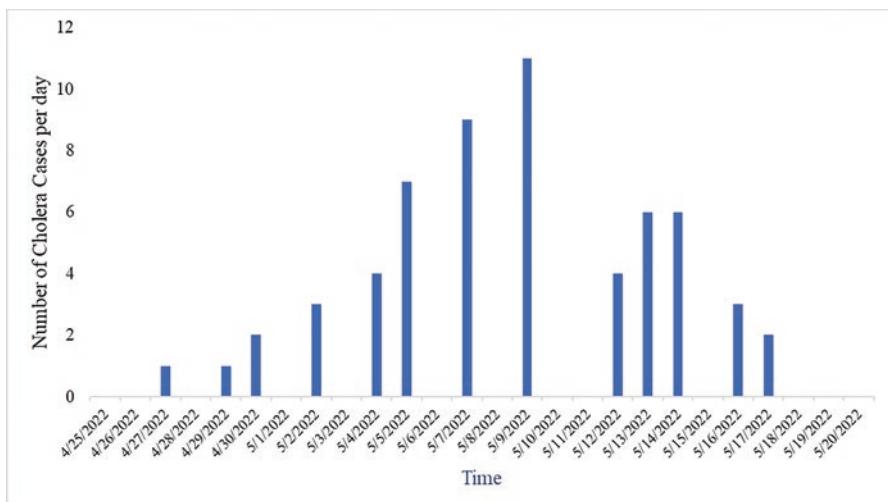
A person can be described for the demographic characteristics (e.g., age, sex, race, ethnicity, income, and occupation). Other characteristics need to be explored according to the situation. For example, in the case of hepatitis B virus outbreak, the outbreak may be concentrated among the intravenous drug users [8].

### 5.6.4 Epidemic Types

Based on the number of cases obtained during the epidemic and the histogram plotted for the cases over time, the epidemic curve shows the onset of the epidemic and it indicates whether it is a sharp or a gradual increase in cases. Similarly, by looking at the curve toward the end of the epidemic, one can determine whether there is a sharp or a gradual fall of the curve. It can be suggested from the nature of an epidemic curve to suggest whether it is a common source or a propagated source of origin of the epidemic.

In the case of common source origin, the epidemic spreads from a contaminated source (i.e., food, water, or drink). Examples: a food-borne outbreak (such as salmonella infection) is spread through a contaminated food; a contaminated community water supply can spread cholera. The epidemic curve of a common source origin commonly shows a tight temporal clustering, with a sharp upslope and a gentle down slope (Fig. 12.3).

In the case of propagated source, the spread of the epidemic takes place either directly from person-to-person or though vectors/vehicles. For example, outbreaks of hepatitis B virus infections/shigellosis. The epidemic curve of such outbreak



**Fig. 12.3** Epidemic curve of common source origin

shows a relatively gentle upslope and then steeper tail. Sometimes there is a second peak, but it is less prominent (Fig. 12.4).

## 5.7 Develop Hypotheses

The next step is to develop a hypothesis. The descriptive epidemiology provides information in developing the hypothesis. Example of a hypothesis: “Date palm sap is associated with *Nipah* virus outbreak in rural Bangladesh.”

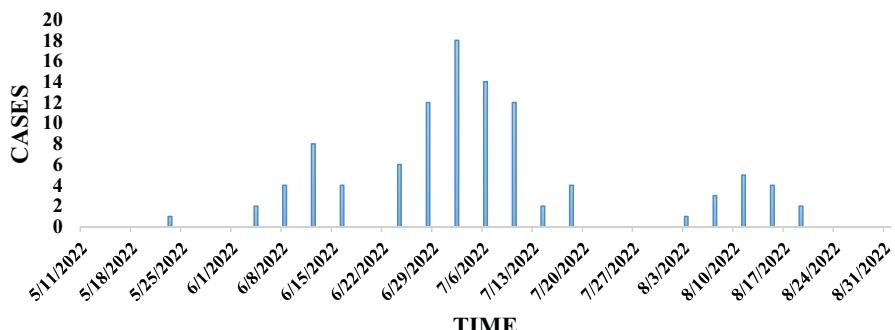
## 5.8 Evaluate Hypotheses

The hypothesis can be evaluated by further epidemiologic investigation, in two ways: (i) comparing the hypothesis with established facts and (ii) quantifying the relationships and assessing the role of “chance” using analytical epidemiology.

The first method can be used where all the evidence (epidemiological, clinical, environmental, and laboratory) supports the hypothesis. Then further investigations are done to test the hypothesis. The second method is utilized where the evidence is not straight forward, and the hypothesis needs to be tested by analytical epidemiology. We can follow different epidemiological study methods to test the hypothesis.

### 5.8.1 Retrospective Cohort Study

In a well-defined, small population, a retrospective cohort study should be the method of choice. Here, the investigators contact each member of the population (e.g., wedding guests). The investigators try to determine the possible exposure sources (e.g., food and beverages consumed by each guest) and take the clinical



**Fig. 12.4** Epidemic curve of propagated source origin

history after the food/beverage consumption. Usually the attack rate (risk) in both exposed and non-exposed groups are calculated, and a risk ratio is calculated.

$$\text{Attack rate (risk)} = \text{total number of cases} / \text{population at risk (exposed)}$$

$$\text{Risk ratio} = \frac{\text{risk among the exposed population}}{\text{risk among the unexposed population.}}$$

### 5.8.2 Case–Control Study

When the population is well-defined and the exposure information before the illness is available, then a case–control study can be used. In the case–control study, a group of cases (people with disease) is selected and a similar group of people but apparently healthy or free from disease of interest (control) is selected. *Odds ratio* is used to measure the strength of association. A case–control study linked date palm sap with *Nipah* virus outbreak in Bangladesh back in 2008 [9].

### 5.8.3 Source of Bias

There is a possibility of existence of bias in the epidemiological analysis. More details on bias are discussed in Chap. 13. Among many biases, the following bias are highlighted here [8]:

- (a) **Recall bias:** The participant's recall may be incorrect due to long incubation period of a disease (e.g., hepatitis A) or due to long time difference between illness onset and the interview. Recall bias may result from an inaccurate exposure recall among the individuals suffering from disease compared to their healthy counterparts [8].
- (b) **Selection bias:** Selection bias occurs in the case–control study when control group is not selected from the population from where the cases had been selected [8].
- (c) **Participation bias:** Participation bias may result in inaccurate estimation of risk. Providing incentives to participate in the study may result in this participation bias. If the incentive is less for the healthy individuals, the participants in the study might not represent the healthy individuals [8].
- (d) **Social desirability bias:** In the case of investigating risk factors which is sensitive or stigmatizing (e.g., drug use, sexual activity, and illicit behaviors), the situation may lead to social desirability bias. The participants may provide inaccurate information about the risk factors [8].

Depending on the results of hypothesis testing, one may re-evaluate, refine, or reconsider a previous hypothesis.

## ***5.9 Compare and Reconcile with Laboratory and/or Environmental Studies***

Laboratory and/or environmental studies are required to compare and confirm the epidemiological findings.

## ***5.10 Implement Control and Prevention Measures***

These measures are simultaneous, ongoing, and not to wait till investigations are over. The initial and immediate measures are somewhat general in nature, as per broad assumption without any concrete knowledge of source/reservoir and route of transmission of the disease. Often such measures take the form of restriction of mass gatherings, temporary disruption of public transport, closure of religious places etc. Its extreme form, such as “lockdown” had been widely employed all over the world during the current pandemic of Covid-19.

Afterward, specific measures are adopted based on the ongoing results of investigations in the form of removing the source/reservoir (agent), interrupting the transmission (environment and vector), and case management with immunization (host). Some specific measures are as follows:

### **5.10.1 Water- and/or Food-Borne Outbreak**

Access to safe drinking water has to be provided including sanitary disposal of human waste. Personal hygiene, especially frequent hand washing with soap and water, is essential. Adopting safe practices in food preparation and handling in eateries are of prime importance.

### **5.10.2 Vector-Borne Outbreak**

Usual vector-control measures need to be followed meticulously. Methods may include space spraying, source reduction, environmental control, and personal protective measures.

### **5.10.3 Respiratory Outbreak**

Much lessons have been learned all over the world during this pandemic of Covid-19. Proper use of mask, maintaining interpersonal distance, and repeated hand washings with soap water/sanitizer have a definite role to prevent the spread of the virus.

#### **5.10.4 Vaccine Preventable Disease Outbreak**

Timely effective vaccination is the best answer when vaccines are available. Its role has also been observed during the pandemic of Covid-19. Proper arrangements of vaccines, syringes and needle, maintenance of appropriate cold chain system, and necessary manpower are essential. For some outbreaks, there is definite role of “Ring immunization.” In ring immunization, the contacts of all susceptible person or confirmed patients are vaccinated. This strategy was used for the control of smallpox.

#### **5.11 *Initiate or Maintain Surveillance***

The control and preventive measures need to be continued. If there is no surveillance system, an active surveillance system needs to be initiated. If the surveillance is already in place, it should be enhanced and continued. The surveillance system will provide information on the effectiveness of the control and prevention measures. An effective surveillance program is essential to detecting disease outbreaks quickly before they spread, cost lives, and become difficult to control. An active surveillance system can help us detect an unknown diseases and take appropriate control measures.

#### **5.12 *Disseminate Findings***

At the end, the findings need to be communicated to the appropriate people (i.e., health department, local public health professionals, and the public). It can be communicated by oral brief or by written report. The final report should be widely circulated to all public health managers at all levels through bulletins, newsletters, web postings, etc. for their appraisal and application in their respective field of work. A copy of this report should also be sent to state/national/international authorities as necessary. The usual reporting schedule in an outbreak comprises a preliminary report by nodal or local public health officer (in the form of first information report—FIR), followed by interim report and the final report as described above. Daily situation update must be provided by local public health authority emphasizing trends in cases and deaths, containment measures, buffer stocks (drugs, equipment, vaccines, etc.), logistics, communication, vehicles, community involvement, and nature of media response. The report should contain the followings:

### **5.12.1 Background**

It includes the geographical and climatic situations of the locality including its demographic and socioeconomic condition. The status of the local health and surveillance system with the provision of any early warning signal should be mentioned. Of course, the “normal” disease pattern in the area should be noted.

### **5.12.2 Previous History of Similar Epidemic**

Previous occurrence of similar epidemic or related disease(s) in the same or neighboring areas needs to be mentioned. The identification of first case(s) of the present outbreak (called index case) should be highlighted.

### **5.12.3 Investigation Methods**

A detailed and meticulous description of the methods of investigations should be mentioned—e.g., case definition, questionnaire(s) used to collect data, survey techniques including surveillance (both retrospective and prospective studies), collection and analysis of laboratory specimens, etc. Information regarding relevant environmental and ecological factors should be included.

### **5.12.4 Data Analysis and Results**

Detailed discussion should be made regarding the clinical findings like symptoms, signs, course of the disease, differential diagnosis, and complications (if any). It is to be followed by a meticulous description of epidemiological analysis, such as frequency of cases, distribution according to time, place, person; mode of transmission including route(s) of entry and exit of pathogen; and factors associated with the occurrence and transmission. A critical note should be there regarding laboratory findings like isolation of agent, drug sensitivity pattern, and serological findings.

### **5.12.5 Interpretation of Results**

A comprehensive picture of outbreak *hypotheses* be presented, including how these were formulated, tested, and retested to reach the final conclusion.

### 5.12.6 Control Measures

The simultaneous and ongoing control measures be reported. A note for the justification of the measures followed by the strategies and implementation of such measures should be described in detail. Results of such control measures including constraints in implementation are valuable information for the future attempts to intervene such an epidemic. Finally, the effectiveness of control measures, especially a cost-effectiveness analysis, should be a part of the report.

### 5.12.7 Future Preventive Measures

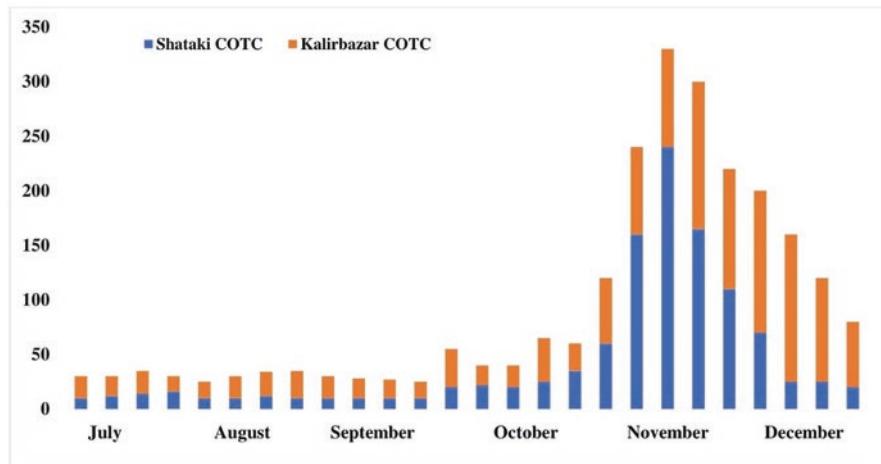
Measures to prevent such outbreak in the future should be an important part of report in the form of suggestions and recommendations.

## 6 “End” of an Outbreak

Declaration of “Outbreak is Over” usually done by the district surveillance officer (DSO), following the standard criterion of “No new cases during two incubation periods since the onset of last case (after *careful search* to make sure no cases are missed).”

### 6.1 *Community-Operated Treatment Centers to Control Diarrheal Epidemic in Rural Bangladesh: A Case Study*

Cholera is endemic in Bangladesh. The International Centre for Diarrheal Disease Research, Bangladesh (ICDDR,B) operates to prevent and control cholera and other diarrheal diseases in collaboration with the government of Bangladesh. At *Shataki* sub-station of *Matlab* Hospital in rural Bangladesh, a makeshift hospital called community-operated treatment center (COTC) was established at the peak time of a cholera epidemic in the region. The COTC was proven effective to manage the patient load, mostly by the help of the community people. The Shataki COTC model was replicated in *Kalirbazar* region of *Matlab* Hospital in June 1982. In October 1982, the catchment area of this COTC was expanded, following a cholera outbreak in *Matlab* and adjacent sub-districts. Between October and December 1982, these two COTCs in Shataki and Kalirbazar treated over 2000 cholera patients (Fig. 12.5). The COTCs averted approximately 820–1000 deaths, proving an effective role COTC in the management of diarrheal epidemics in resource-limited rural settings [10].



**Fig. 12.5** Number of patients treated by week, at the Shataki and Kalirbazar Community-Operated Treatment Centers (COTC), between 1 July and 31 December 1982 [10]

## 7 Conclusion

Outbreak investigations warrant a combination of field work and epidemiological investigation. A new disease outbreak due to monkeypox is currently affecting many parts of the world. It is possible that a new and emerging disease outbreak such as the COVID-19 pandemic has the potential to spread across the globe, disrupt social life, and threaten the global economy. That is why the outbreaks need to be controlled in the quickest possible time. Sometimes, in addition to health authorities, a holistic approach may be necessary to combat the outbreak.

## 8 Problem Solving

In September 2022, 15 new cases of tuberculosis and 19 new cases of West Nile virus infection were reported to a county health department. As a county health epidemiologist, it is your role to decide whether this is an outbreak or not. What additional information do you need?

### Answer

Tuberculosis is not a seasonal disease. The number of cases in the month of September can be compared to: (1) number of cases in the prior few months and (2) number of cases in September of the previous years.

West Nile virus infection is a seasonal disease and August–October is the peak time for the viral infection. The number of cases of West Nile infection in the month of September ( $n = 19$ ) should be compared with the number of West Nile virus

infection cases in the month of September of the previous few years. If the number of cases now exceeds the expected number of the same time in previous years, the current number may constitute an epidemic. However, further investigations are warranted to confirm the epidemic.

## 9 Further Practice

1. Which of the following is a part of an outbreak investigation?
  - (a) Generating a hypothesis
  - (b) Communicate findings
  - (c) Initiate or maintain surveillance
  - (d) All of the above
2. An epidemic curve is used to describe the epidemic in terms of what?
  - (a) Place
  - (b) Time
  - (c) Person
  - (d) None of the above
3. In a water- and or food-borne outbreak, which is **not** an appropriate control measure?
  - (a) Wearing mask
  - (b) Maintaining personal hygiene
  - (c) Washing hand before eating and after using washroom
  - (d) Drinking clean water
4. What is the source of bias in the epidemiological analysis during an outbreak investigation?
  - (a) Social desirability bias
  - (b) Participation bias
  - (c) Recall bias
  - (d) All of the above
5. A case definition during an outbreak investigation should specify
  - (a) Clinical features
  - (b) Place
  - (c) Person
  - (d) All of the above
6. Define the following terms in your own word.
  - (a) Outbreak
  - (b) Epidemic

- (c) Endemic
  - (d) Pandemic
7. What constitutes a Rapid Response Team?
  8. What are the characteristics of a common source epidemic?
  9. What are the characteristics of a propagated source epidemic?
  10. Define “end of an outbreak” in your own word.

### Answer Keys

1. d
2. d
3. a
4. d
5. d
6. see Sect. 2
7. see Sect. 4.1
8. see Sect. 5.6.4
9. see Sect. 5.6.4
10. see Sect. 6.

### References

1. Centers for Disease Control and Prevention. Principles of epidemiology, Lesson 1, Section 11. [cited 1 Sep 2022]. Available: <https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section11.html>.
2. Centers for Disease Control and Prevention. Principles of epidemiology, Lesson 3, Section 2. [cited 11 Oct 2022]. Available: <https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section2.html>.
3. Encyclopedia Britannica. Case fatality rate. [cited 11 Oct 2022]. Available: <https://www.britannica.com/science/case-fatality-rate>.
4. Centers for Disease Control and Prevention. Principles of epidemiology, Lesson 1, Section 8. [cited 11 Oct 2022]. Available: <https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section8.html>.
5. Integrated Disease Surveillance Programme. Training manual for state & district surveillance officers; Outbreak Investigation, Response & Control. [cited 23 May 2022]. Available: [https://idsp.nic.in/WriteReadData/OldSite/2WkDSOSept08/Resources\\_files/DistrictSurvMan/Module8.pdf](https://idsp.nic.in/WriteReadData/OldSite/2WkDSOSept08/Resources_files/DistrictSurvMan/Module8.pdf).
6. Centers for Disease Control and Prevention. Principles of epidemiology, Lesson 6, Section 2. [cited 11 Sep 2022]. Available: <https://www.cdc.gov/csels/dsepd/ss1978/lesson6/section2.html>.
7. Centers for Disease Control and Prevention. Public health dispatch: outbreak of listeriosis—Northeastern United States, 2002. Morbid Mortal Wkly Rep. 2002;51(42):950–1.
8. Tam C, Haas W. Outbreak investigations. In: Infectious disease epidemiology; 2016. p. 35–52. <https://doi.org/10.1093/MED/9780198719830.003.0003>.
9. Rahman MA, Hossain MJ, Sultana S, Homaira N, Khan SU, Rahman M, et al. Date palm sap linked to Nipah virus outbreak in Bangladesh, 2008. Vector Borne Zoonotic Dis. 2012;12:65–72. <https://doi.org/10.1089/VBZ.2011.0656>.
10. Baqui AH, Yunus MD, Zaman K. Community-operated treatment centres prevented many cholera deaths. J Diarrhoeal Dis Res. 1984;2:92–8.

# Chapter 13

## Population Projection



Mohammad Mainul Islam and Amal K. Mitra

### Learning Objectives

After completing this chapter, you will be able to:

- Understand the population projection and make the distinction between population projection and population estimate
- Explain how population projections can be useful decision-making tools for policymakers
- Illustrate the role of population projection in public health
- Introduce the common methods used for making population projections and identify the type of outputs required by such projections
- Calculate the appropriate measures of population projection—cohort component and mathematical approaches to population projections

## 1 Introduction

Generally, a population projection demonstrates a picture of the future size and structure of the population by sex and age. This is based on knowledge of past trends and, for the future, on assumptions made for three vital demographic processes: fertility, mortality, and migration. The projections serve as a basis for long-term thinking for overall development. In this regard, this chapter will explain

---

M. M. Islam (✉)

Department of Population Sciences, Faculty of Social Sciences, University of Dhaka,  
Dhaka, Bangladesh  
e-mail: [mainul@du.ac.bd](mailto:mainul@du.ac.bd)

A. K. Mitra

Department of Epidemiology and Biostatistics, Jackson State University, Jackson, MS, USA  
e-mail: [amal.k.mitra@jsums.edu](mailto:amal.k.mitra@jsums.edu)

the conceptualization, conduct, and interpretation of population projection to the extent of the figure of population size for a country or locality. The chapter will address the scopes, types, and measurements of population projection. This will explain the distinction between a population projection and a population estimate and how to use population projection in the planning process to understand the impact of planning activities on population size, composition, and distribution. The conditions under which population projection and estimation are needed will also be described. This will illustrate the role of population projection in public policy and public health, introduce the standard methods used for making population projections, and demonstrate the appropriate measures like cohort components and mathematical approaches to population projections.

## 2 What Is Population Projection? How Do Projections Differ from Estimates?

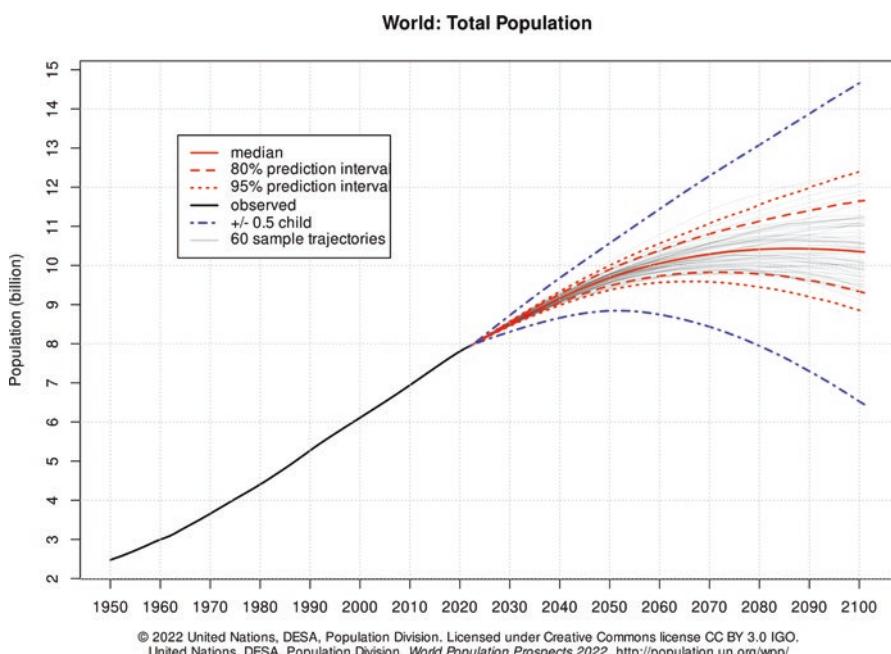
Population projection is a method of computing future changes in population size, by gender and age group, given certain assumptions about future trends in the rates of fertility, mortality, and migration—the key demographic elements which determines the population growth. Population Scientists usually do projections based on previous rates and trends for the future. A number of assumptions are made about scenarios using low, medium, and high rate of changes of the same population in the future [1]. But, population projections are always based on a “conditional” future because the fact is that the changes from the baseline number of population are assumptions and not an absolute truth. Thus, populations can be projected either by extrapolating trends in the growth in the total population or by projecting each age-sex cohort separately using assumptions about age-specific rates of fertility, mortality, and migration. For that, the prediction of a future population is based on the knowledge of the past events in the population, such as fertility rates and mortality rates. Later an adjustment is made for net migration (number of immigrants minus number of emigrants). Projection is usually completed in 5- or 10-year phases, with probabilities of dying across the 5- or 10-year period. Probabilities of death is provided by an appropriate life table analysis. Assumptions of future mortality and fertility are usually required, and models are often used in this process.

There is a distinction between the two terms—projection and estimation. The term “estimation” implies a judgment or assumptions as to the size of the historical or current population and also changes in demographics including age, sex, education level, occupation, etc., whereas “projection” means a judgment or assumption as to the future populations based on its current attributes. The simplest population projections begin with population estimates at two or more time points or with the population size and either birth and death rates or the net reproduction rate [2]. While projections and estimates may appear similar, the two measures have some distinct differences. Estimates are for the recent past, near future, and more often for

intercensal periods, while projections are based on assumptions about future demographic trends. Estimates generally use existing data collected from various sources, while projections make assumptions about what demographic trends will be in the future.

Population projection has many uses. The government uses a projected population to estimate and collect taxes. Planners use projected population to allocate resources to build up infrastructures. Social scientists use the future population to formulate social programs and interventions. Many other uses of projected population structures by age group and gender include planning for the construction of schools, the training of teachers, the recruitment of labor, and the provision of health services. Calculations of the future population for periods vary from 1 to 30 years or sometimes even further ahead. This is especially important for social security planning: short-term assessments are required for budgetary purposes, and projections of a more extended period are warranted to illustrate the broader financial trends [3]. Making projections is most appropriate when vital rates and age structure can be assumed to be constant or if the only thing one knows about the population is its total size. Figure 13.1 is an example of a projected world population.

**Explanation of data in Fig. 13.1** Above chart shows estimates and probabilistic projections of the total population for countries or areas, geographical aggregates,



**Fig. 13.1** Example of UN projected world population, 1950–2100 [4]

and World Bank income groups as defined in the Definition of Regions. The population projections are based on the probabilistic projections of total fertility and life expectancy at birth. These probabilistic projections of total fertility and life expectancy at birth were carried out with a Bayesian Hierarchical Model. The figures display the probabilistic median, and the 80% and 95% prediction intervals of the probabilistic population projections, as well as the (deterministic) high and low variant ( $\pm 0.5$  child).

### 3 Scope and Use of Population Projection

Demography is mainly concerned with answering questions about how populations change over time and how we measure them. In this regard, population projections vary widely in their geographic territories, time limit, and types of outcome measurements used. Spatial dimensions can range from local areas to the entire world. The diversity of the kinds of projections is driven by the variations in users' objectives [5]. Population projections are used to inform policy and planning in a wide range of contexts by national and local governments and private sectors. The information on the future population forecasts varies significantly based on: *geographical details, subject-specific disaggregation, and future duration*. The projected population is the basis for all forms of planning for the future—Involving the social, economic, or business planning. For example, suppliers of any good or service can only plan if they have the idea of the size of the population and potential users to be located. It is the social, economic, and environmental changes, along with government policies, that can influence future demographic trends, especially fertility level, which ensure some uncertainty in population projections. No doubt, population projection is important since it helps the stakeholders, including the government and researchers, make decisions about the future. Policymakers and program planners can undertake steps to make projections more useful for policy and planning purposes. For that, understanding the causes of uncertainty in population projections and the implications of this uncertainty for plans and policies that span different time horizons and target specific population groups are needed. This will contribute to national and international efforts to collect more accurate demographic data, which would lead to more accurate assumptions about fertility, mortality, and migration and better projections; and cooperate with national and international research efforts to develop more accurate projections by supporting organizations that investigate better projection methodologies. It is important to have high-quality data on the population number and projections of the population for policy development, planning, and providing public services. The uses of population projection may be attributed to central and local finance allocation; informing local and national policy; childcare and schools planning; housing and land use planning; health care planning; modeling and projecting health care indicators; weighting surveys; benchmarking other projections and as a control for smaller area projections; looking at the implications of an aging population; and making national and

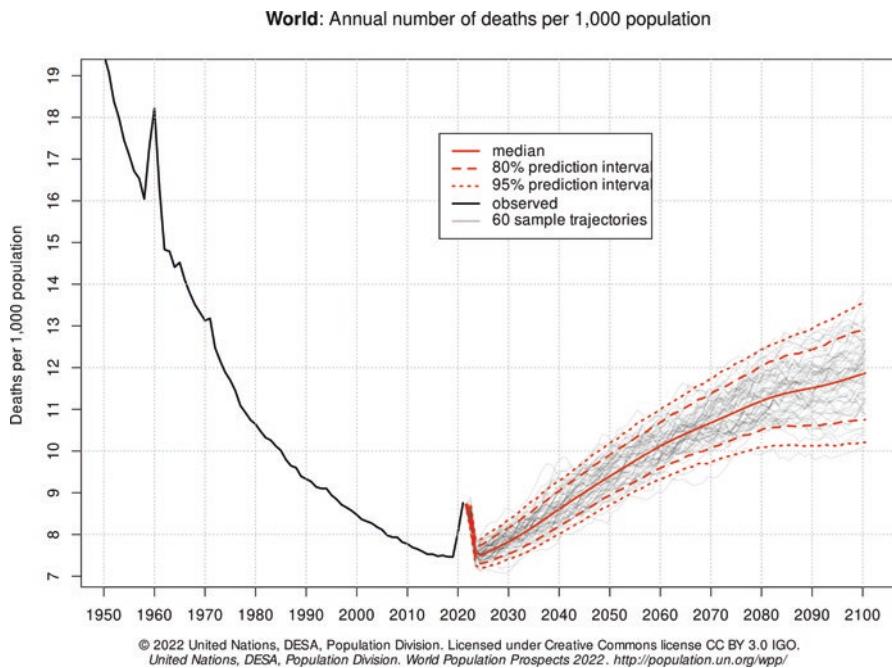
international comparisons. Population projections can alert policymakers to significant trends affecting economic development and help policymakers craft policies that can be adapted for various projection scenarios.

## 4 Public Health Importance of Population Projection

Public health is concerned with the changing nature of diseases in the population. Changes in population characteristics affect the disease pattern. There is a need for population projections in the field of public health. A population's health and healthcare needs cannot be measured or met without the knowledge of its size and characteristics. Demography is concerned with this and with understanding population dynamics—how populations change in response to the interplay between fertility, mortality, and migration. This understanding is a pre-requisite for making projection about future population size and structure which should underlie healthcare planning. The number of births in a population of a country depends on current patterns of family formation and the number of women of reproductive age—a function of past trends in fertility and mortality. Similarly, the number and causes of death are strongly influenced by age structure. For example, prevalence of disabilities, injuries, and mental health issues are increasing with the increase in the life expectancy of the people and the increase in the proportion of elderly people. Information on demographic methods, data sources, and their applications to health and population issues greatly influences public health policies. Thus, an understanding about demographic trends through projection and their implications, and information about the complex inter-relationship between population change and human health are vital [6].

Government planners and policymakers may be concerned with population aging and its potential social and economic impact. They may therefore desire longer-term projections and want to know more about the health status and living arrangements of the older persons [5]. As a demographic process, mortality is incorporated into projections by estimating death rates by age group and sex. Figure 13.2 is a demonstration of projected crude death rates (CDRs) in the world population.

Where mortality is relatively high and the resulting life expectancy at birth is relatively low, changes in mortality play an essential role in future population size. On the contrary, where mortality is already low and life expectancy has increased, mortality has much less effect. Throughout developing countries, infant mortality has declined substantially over the last several decades; the general assumption underlying population projections for all countries is a continued decline in death rates and an increase in life expectancy at birth. Figure 13.3 shows projected life expectancy at birth in the world population.



**Fig. 13.2** Probabilistic projections of crude death rate (CDR) in the world. (Source: United Nations. Department of Economic and Social Affairs, Population Division. World Population Prospects, 2022 [4])

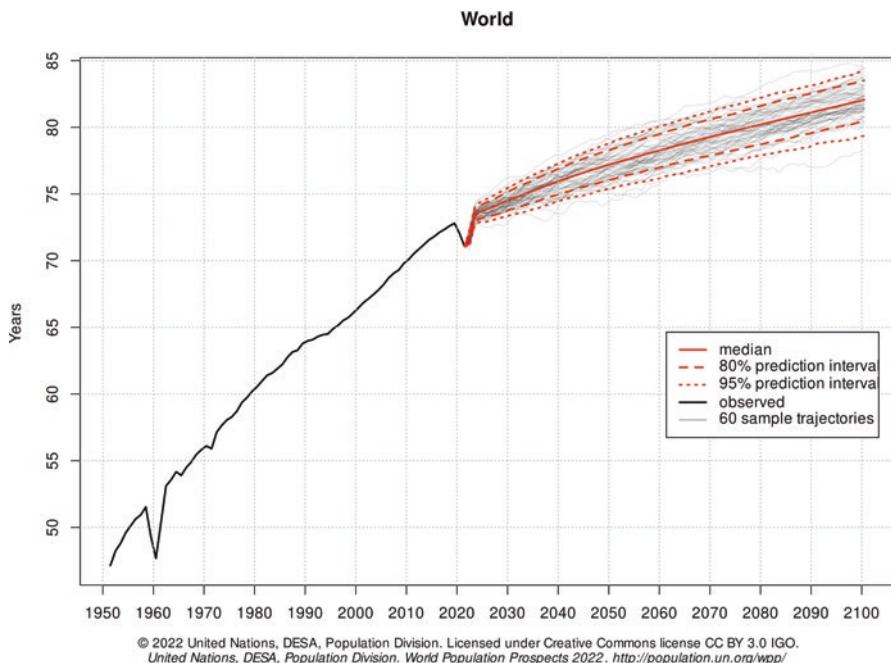
## 5 Measurement and Applications with Examples

Population projections are estimates of the population for future dates. Projections illustrate possible courses of population change based on assumptions about future births, deaths, net international migration, and domestic migration. Several methods are used for the calculation of population projection:

- **Mathematical method of projection:** *Arithmetic projection; geometric projection; exponential projection*
- **Component method of projection:** *Cohort component method*

### 5.1 Arithmetic Projection

Arithmetic projection has generally been used to show the population growth in fixed number through successive equal period of time based on the observation of the number of population increase at the base period. This is also called **a linear growth**.



**Fig. 13.3** Probabilistic Projections on Life Expectancy of Both Sexes in the World. (Source: United Nations. Department of Economic and Social Affairs, Population Division. World Population Prospects, 2022 [4])

For example, the population of *Country X* has been observed to increase on an average by 35,000 every 10 years since 1960, with an initial population of 200,000 in 1960. What would be the expected population in 2030?

**Calculations** Year 2030 – Year 1960 = 70 years (seven decades).

$$35,000 \times 7 = 245,000 \text{ increase.}$$

$$\text{The expected population in 2030} = 200,000 + 245,000 = 445,000.$$

## 5.2 Geometric Projection

Geometric projection relies on the population changes in terms of percentage changes, rather than numerical changes. It is called a **relative growth rate**, which is expressed as percentage. The geometric projection method has been much more popular than the arithmetic projection method.

**Example** The population size of Country *Y* was 340,000 in 2000. This is 60% greater than the population in 1970. What would be the projected population in 2030?

**Calculations** By using a **simple trend projection**, the population is expected to increase 60% in next 30 years from Year 2000.

$$340,000 \times 60\% = 204,000.$$

$$\text{The expected population in 2030} = 340,000 + 204,000 = 544,000.$$

### 5.3 Exponential Projection

In some countries, especially in several developing countries such as Bangladesh, China, and India, population grows exponentially. A quantity grows exponentially if it grows by a constant factor or a rate for each unit of time.

**Example** The City population is growing at a rate 1.6% per year. The initial population of the City in 2020 is 130,000. Calculate the City's population over the next 5 years.

#### Calculations

$$P_0 = 130,000$$

The relative growth rate = 1.6%. This means an additional 1.6% will be added on total 100% of the population that already exists in the preceding year.

$$100 + 1.6 = 101.6\%$$

Population in 2021 (in 1 year):

$$130,000 (1.016)^1 = 132,080$$

Population in 2022 (in 2 years):

$$132,080 (1.016) = 134,193.28$$

$$\text{Or, } 130,000 (1.016)^2 = 134,193.28$$

Population in 2023 (in 3 years):

$$134,193.28 (1.016) = 136,340.37$$

$$\text{Or, } 130,000 (1.016)^3 = 136,340.37$$

Population in 2024 (in 4 years):

$$136,340.37 (1.016) = 138,521.82$$

$$\text{Or, } 130,000 (1.016)^4 = 138,521.82$$

Population in 2025 (in 5 years):

$$138,521.82 (1.016) = 140,738.17$$

$$\text{Or, } 130,000 (1.016)^5 = 140,738.17 \text{ or approximately } 140,738$$

From the above calculations, we can deduct an equation as follows:

$$P(t) = P_0 (1+r)^t$$

where  $P(t)$  = population after a given time (in year)

$P_0$  = Baseline population (or current population)

$r$  = population growth rate

$t$  = time in year

### 5.3.1 Shrinking Population

In the event that the population is shrinking, instead of growing, you would subtract the population growth rate from 100.

**Example** The population of California was 39,303,157 in 2021. The current population (2022) of California declined by 0.3% in 1 year to 39,185,605. If the rate of decline remains the same, what would be the population in 2025?

#### Calculations

$$P_0 = 39,185,605 \text{ in 2022}$$

$$r = 100 - 0.3 = 99.7\% \text{ or } 0.997$$

$$t = 3 \text{ (population after 3 years)}$$

$$\begin{aligned} P(t) &= P_0 (1-r)^t \\ P(t) &= 39,185,605 (1-0.3)^3 \\ &= 39,185,605 (0.997)^3 \\ &= 38,833,992 \end{aligned}$$

## 5.4 Cohort-Component Method of Population Projection (CCP) [7]

The procedure of making cohort-component population projection was developed by Whelpton in the 1930s. In the cohort component method of projection (CCP), the components of population change, such as fertility, mortality, and net migration, are projected separately for each birth cohort (persons born in a given year) [7]. Net migration is the difference between the number of immigrants and the number of emigrants. In this calculation, the base population is advanced each year by using projected survival rates and net international migration. Also, it makes the use of the fact that every year of time that passes, every member of a population becomes a year older. Thus, the CCP algorithm provides a complete account of past population dynamics, including size, growth rates, and changing age-sex composition. Another advantage of using CCP is that given assumptions about future inputs, future population dynamics follow with certainty. However, there are some limitations in using

CCP. First, CCP is highly dependent on reliable birth, death, and migration data, which may be difficult to collect the information to apply this method. Second, it assumes that survival and birth rates and estimates of net migration will remain the same throughout the projection period. Third, CCP does not consider the nondemographic factors that influence population growth or decline.

Despite these limitations, CCP is the most widely used method since it provides information on the potential growth or decline of a population by age and sex.

## 6 Problem Solving

In this section, we will use the formula shown in Sect. 6.3 to calculate populations of several countries by using the exponential projection method.

### 6.1 Problem 1

The current population of Mississippi is estimated to be 2,961,279. Calculate the time (in years) needed to get the population size of 3.5 million. The current rate of population growth in Mississippi is 0.30%. Assume that the rate of population growth will remain unchanged.

Formula

$$P(t) = P_0(1+r)^t$$

#### Calculations

$$r = 0.30\% = 0.003$$

$$P_0 = 2,961,279$$

$$P_t = 3,500,000$$

$$3,500,000 = 2,961,279 (1 + 0.003)^t$$

$$\text{Or, } 2,961,279 (1 + 0.003)^t = 3,500,000 \text{ (changed the side of the equation)}$$

$$\text{Or, } (1 + 0.003)^t = 3,500,000 / 2,961,279$$

$$\text{Or, } (1.003)^t = 1.1819$$

Now, use log of both sides.

$$\text{Log } (1.003)^t = \text{Log } (1.1819)$$

$$\text{Or, } t * \text{Log } (1.003) = \text{Log } (1.1819)$$

$$\text{Or, } t = \text{Log } (1.1819) / \text{Log } (1.003)$$

$$\text{Or, } t = 0.0726 / 0.0013$$

$$t = 55.8 \text{ years}$$

**Conclusions** It will take about 56 years to get the population of 3.5 million in Mississippi, provided the rate of population growth per year remains the same.

## 6.2 Problem 2

Bangladesh is a highly populous country, having 164,689,383 people, and a population growth rate of 0.98%. How many years will it take to get the population size to 200 million?

$$r = 0.98\% = 0.0098$$

$$P_0 = 164,689,383$$

$$P_t = 200,000,000$$

Formula

$$P(t) = P_0(1-r)^t$$

### Calculations

$$200,000,000 = 164,689,383 * (1 + 0.0098)^t$$

$$\text{Or, } (1.0098)^t = 200,000,000 / 164,689,383$$

$$\text{Or, } (1.0098)^t = 1.2144$$

Now, use log of both sides of the equation.

$$\text{Or, } t * \log(1.0098) = \log(1.2144)$$

$$\text{Or, } t = \log(1.2144) / \log(1.0098)$$

$$t = 0.0844 / 0.00424$$

$$t = 19.9 \text{ years}$$

**Conclusions** It will take approximately 20 years to have 200 million population in Bangladesh.

## 6.3 Problem 3

The population of Canada is 38,415,364 as of July 2022. The rate of population growth of the country is 0.78% per year. What would be population size of Canada after 10 years?

$$r = 0.78\% = 0.0078$$

$$t = 10$$

$$P_t = 38,415,364 (1 + 0.0078)^{10}$$

$$\text{Or, } P_t = 38,415,364 * (1.0078)^{10}$$

$$\text{Or, } P_t = 41,519,153$$

**Conclusions** After 10 years, the population of Canada would be approximately 41,519,153.

## 6.4 Problem 4

According to the United Nation's estimate, India's current population is 1,406,631,776 and the population growth rate is about 0.97%. Find out how long it will take to increase the population to two billion.

$$r = 0.97\% = 0.0097$$

$$P_0 = 1,406,631,776$$

$$P_t = 2,000,000,000$$

Formula

$$P(t) = P_0(1-r)^t$$

### Calculations

$$2,000,000,000 = 1,406,631,776 * (1 + 0.0097)^t$$

$$\text{Or, } 1,406,631,776 * (1 + 0.0097)^t = 2,000,000,000$$

$$\text{Or, } (1 + 0.0097)^t = 2,000,000,000 / 1,406,631,776$$

$$\text{Or, } (1.0097)^t = 1.4218$$

Now, use log of both sides of the equation.

$$\log(1.0097)^n = \log(1.4218)$$

$$\text{Or, } n * \log(1.0097) = \log(1.4218)$$

$$\text{Or, } n = \log(1.4218) / \log(1.0097)$$

$$n = 0.1528 / 0.00419$$

$$n = 36.5 \text{ years}$$

**Conclusions** It will take approximately 36.5 years to get two billion population in India.

## 7 Further Practice

1. What is the difference between projection and estimation?
2. How does population project help public health?
3. What are the components of population change?
  - (a) Marriage
  - (b) Birth
  - (c) Death
  - (d) Birth, death, and migration
4. The following characteristics of a geometric progression true, except
  - (a) Geometric projection relies on the population changes in terms of percentage
  - (b) Geometric changes mean numerical changes
  - (c) Geometric projection is also called a relative growth rate
  - (d) Geometric projection method has been much more popular than the arithmetic projection method.

5. Which is the most widely used method of population projection?
  - (a) Cohort component method
  - (b) Arithmetic projection
  - (c) Geometric projection
  - (d) Exponential Projection
6. When would be the cohort component projection method be used?
  - (a) When population projections by age and sex are needed for 5 years
  - (b) When population projections by age and sex are necessary for 10 years
  - (c) When population projections by age and sex are required for more extended periods
  - (d) All of the above
7. Cohort-component method of population projection (CCP) provides information on the potential growth or decline of a population by age and sex.  
True / False
8. CCP considers the nondemographic factors that influence population growth or decline.  
True / False.
9. What are the weaknesses of the cohort-component method?
10. What do you mean by the term “Exponential” growth?

### Answer Keys

1. “Estimation” implies a judgment or guess as to the size or attributes of a historical or present population; “projection” means a judgment or guess as to its future direction.
2. Public health is concerned with the changing nature of diseases in the population. Changes in population characteristics affect the disease pattern. Population projections can alert policymakers to significant trends affecting economic development and help policymakers craft policies that can be adapted for various projection scenarios.
3. (d)
4. (b)
5. (a)
6. (d)
7. True
8. False
9. (a) It is highly dependent on reliable birth, death, and migration data, which may be difficult to collect the information to apply this method.  
(b) It assumes that survival and birth rates and estimates of net migration will remain the same throughout the projection period.  
(c) It does not consider the nondemographic factors that influence population growth or decline.
10. It is the rate of population growth that occurs exponentially; that means it grows by a constant factor or a rate for each unit of time.

## References

1. Population Reference Bureau. Available at: u (PRB). Understanding and using population projection, policy brief. Washington, DC; 2001. Available at: <https://www.prb.org/resources/understanding-and-using-population-projections/>. Accessed 1 Nov 2022.
2. Smith DP. Chapter 8: population projection and population matrices. In: Formal demography. New York: Plenum Press; 1992. p. 256.
3. Cox PR. Demography. 5th ed. Cambridge: Cambridge University Press; 1976. p. 153.
4. United Nations. DESA, Population Division. World population prospects 2022. 2022. Available at: <https://population.un.org/wpp/Graphs/Probabilistic/POP/TOT/900>. Accessed 10 Nov 2022.
5. O'Neill BC, Balk D, Brickman M, Ezra M. A guide to global population projections. Demogr Res. 2001;4:203–88. <https://doi.org/10.4054/DemRes.2001.4.8>.
6. Grundy E, Murphy M. Demography and public health. In: Detels R, et al., editors. Oxford textbook of global public health. 6th ed. Oxford Textbook (Oxford Academic); 2015. <https://doi.org/10.1093/med/9780199661756.003.0126>. Accessed 15 Oct 2022.
7. The United Nations Population Fund (UNFPA). Cohort-component methods of projection. In: Population projections: concepts and methods. Available at: [http://papp.iussp.org/sessions/papp101\\_s10/PAPP101\\_s10\\_070\\_010.html#4](http://papp.iussp.org/sessions/papp101_s10/PAPP101_s10_070_010.html#4). Accessed 10 Nov 2022.

# Chapter 14

## Geospatial Applications in Epidemiology: Location, Location, Location



Stephen Scroggins

### Learning Objectives

After completing this chapter, you will be able to:

- Define spatial epidemiology and its role in epidemiology and public health
- Describe geographic themes in the context of spatial epidemiology
- Describe spatial autocorrelation
- Illustrate appropriate and common visualization methods used in spatial epidemiology
- Interpret typical hypothesis testing related to spatial epidemiology

## 1 Introduction

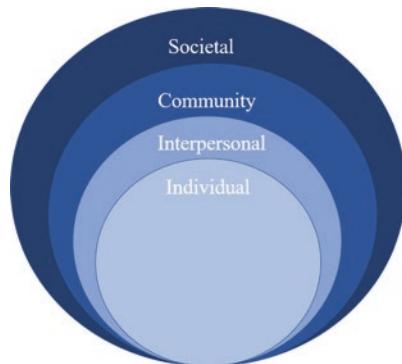
Geographic location has played an integral role throughout the history of epidemiology. Location was a focus of prevention efforts during the fourteenth-century bubonic plague pandemic as cities near ports implemented quarantine requirements [1]. Over 500 years later, John Snow used maps to trace the source of cholera outbreaks in 1854 in London [2]. The early twentieth century gave rise to strategically located tuberculosis sanitariums in dry climate areas [3]. From ancient leprosaria or locations where people with leprosy were isolated to modern COVID-19 hotspot maps, location is intrinsically associated with health [4, 5]. Unmistakably, an individual's health is influenced by their location. While this resolution may seem obvious, its reach is overarching. Considering the socioecological model, the hierarchical framework is used to illustrate varying levels of interacting influence a person may experience, as shown in Fig. 14.1.

---

S. Scroggins (✉)

Saint Louis University, College for Public Health and Social Justice, Saint Louis, MO, USA  
e-mail: [steve.scroggins@slu.edu](mailto:steve.scroggins@slu.edu)

**Fig. 14.1** The socioecological model



**Table 14.1** Levels of the socioecological model and the varying levels of locational influence

Level	Definition	Locational influence
Societal	Macrollevel systems, often implemented, regulated, and enforced, explicitly or not, by state or national agencies	Nation, state, or county of residence
Community	The built environment surrounding an individual, which they may or may not directly interact with	City, town, or neighborhood of residence
Interpersonal	Relationships and social support networks	Location where a person spends time with others (e.g., home, work, activity/leisure spaces)
Individual	Personal beliefs, attitudes, and knowledge	A location where a person spends anytime

At each level, location plays a significant role, as detailed in Table 14.1.

From Table 14.1, we see that the hierarchical scheme of the socioecological model can be translated to a collection of places that people occupy. Each location we occupy has varying, sometimes conflicting, rules, regulations, and norms that influence our health. While these influential interactions are elucidated by spatial epidemiology, spatial epidemiology also goes beyond these first-order associations.

This chapter will discuss the principles and common methodologies used in spatial epidemiology through the context of both research and practice.

## 2 The Distinction of Spatial Epidemiology

Directly or indirectly, location is a fundamental part of most epidemiological practices. The distinction of spatial epidemiology is its focus on variation across locations. To better understand this concept, we can apply Tobler's First Law of Geography: "everything is related to everything else, but near things are more

**Fig. 14.2** Example of autocorrelation among varying shades of blue



related than distant things” [6]. Within the context of epidemiology, we could infer this to mean that the prevalence and incidence of health outcomes do not randomly occur at specific locations but are correlated across varying locations or autocorrelated. Suppose relative-risk ratio (RRR) or disease prevalence calculations are the core of classic epidemiology. In that case, autocorrelation may very well be the core of spatial epidemiology, where consideration is given to a central location and the surrounding locations.

Autocorrelation, like the name would suggest, is the phenomenon of an occurrence being correlated with itself at lagged intervals. A simple example of autocorrelation is illustrated in Fig. 14.2. In Fig. 14.2, each square is filled with various shades of blue, with the top left square completely filled with the darkest shade and the bottom right square completely filled with the lightest shade. As we move farther away from the top left square, each proximal square contains less of the darker shade of blue. In this example, the shade of blue is likely autocorrelated among the squares. Now imagine that each square represents a location (e.g., state, county, and home) and the shade intensity represents the prevalence of some health outcome! A more detailed and statistical explanation of autocorrelation and corresponding hypothesis tests is further discussed in Sect. 3.

While autocorrelation is a key concept of spatial epidemiology, it certainly is not unique to only spatial epidemiology. In addition to spatial epidemiology, autocorrelation of an occurrence or event is common in physics, biology, computer science, and psychological studies [7–10].

## 2.1 The Intersection of Spatial Epidemiology and Geography

Before diving into the deep end of spatial epidemiology, it is first helpful to define common terms related to the science of geography. In the landmark text, “Guidelines for Geographic Education”, five geographical themes are outlined and are helpful

**Table 14.2** The five geographical themes and respective epidemiological examples

Theme	Definition	Epidemiological example
1. Location	The physical space that something occupies on earth	Coordinates of substance use treatment centers
2. Place	A location's characteristics and features	Locations that sell tobacco products
3. Human–environment interaction	The relationships that people have with locations they occupy	Patients experiencing different health outcomes at different hospitals
4. Movement	Travel between locations	Routes people take when they seek medical care
5. Region	Areas with distinctive geographic, demographic, or political characteristics	Dietary differences between urban and rural regions

when discussing spatial epidemiology [11]. These components are detailed in Table 14.2 and discussed below.

If you were to take all the words in this chapter and create a list based on how frequently they appeared, the word “location” would likely be at the top. Location can mean something very broad or specific, depending on the context. So far in this chapter, we have reviewed how location is a fundamental concept to most of the epidemiology and that investigating variation across locations makes spatial epidemiology unique. But what is location and how is it measured?

In spatial epidemiology, location is often measured in one of two distinct manners: (1) points or (2) polygons. As the name would suggest, point locations refer to a precise location. To fully grasp a point location, imagine taking a black dot, like the period at the end of this sentence, and placing it on a blank sphere. How would you describe the black dot’s position? After some trial and error, you would probably come to the same conclusion Descartes did in the seventeenth century and invent a coordinate system. However, a two-dimensional cartesian system, like Descartes developed, would not work since the black dot is on a sphere. We need angles! To achieve this, point locations are usually recorded in latitude and longitude, or the measurement of angles from the sphere’s center to the sphere’s surface. Applying a geographic coordinate system, of which many exist, allows us to identify any point on Earth. Using point locations in spatial epidemiology allows us to make inferences and predictions based on distance since we typically have multiple observations of varying precise measurements. This is typically achieved using a point-process model or point-pattern analysis. These models are especially useful when predicting infectious disease transmission like Ebola or Influenza [12, 13] as they can help inform where the probability of transmission is higher or where prevention efforts may need to be focused.

In addition to points, locations can also be represented by polygons. Polygons are two-dimensional shapes enclosed with contiguous lines representing a border. If you picture an outline of the USA, you are likely picturing a polygon. Now picture all the US states; you are probably picturing a collection of polygons. Spatial polygon data make up the hallmark choropleth maps we often use in spatial

epidemiology and will be discussed later in this chapter. While point location data is precise, polygon locations can also represent a surprisingly wide gradient of scale. For example, polygon data could be a collection of countries and nations or a collection of homes or dwellings. Similar to point locations, polygons are also represented with a geographic coordinate system. Though one latitude and one longitude value can reveal the location of a single point, a polygon location typically comprises multiple latitude and longitude coordinates. While polygon locations have the benefit of often being standardized with administrative borders (e.g., a school district, county, state, or nation), the distance between polygon locations can be difficult to estimate due to the loss of precision. Conversely, while point locations may be more geographically precise, point location data may be intrusive, especially in public health research where topics can be sensitive or protected.

The distinction between location, point or polygon, and place can be difficult to decipher. When does a location become a “place,” and are they different? In the context of spatial epidemiology, a place is what a location contains. For example, you might be given a longitude and latitude, the coordinates of a point location. You look up these coordinates and see a fitness center at that location. A place! This distinction, though slight, is crucial in spatial epidemiology.

Perhaps more than other geographic themes, human–environment interaction is the most straightforward. Epidemiological studies, including spatial studies, have long examined how environmental factors influence human health. A 2005 review investigating participation in physical activity found that the environmental aesthetics of activity facilities were more strongly associated with the facility’s safety or the environment’s weather [14]. Other studies have found significant associations between adverse health behaviors and the frequency of retail stores that sell alcohol and unhealthy foods [15, 16]. Overall, these and similar studies continue to show that environment influences health behavior. More recently, spatial epidemiological studies have also been investigating the inverse of this relationship. Studies show that population growth, agricultural practices, and urban development are fastening the effects of global warming [17–19].

While interaction with the environment is an epidemiological mainstay, the environments in which people interact are not static, that is to say, people move around. This movement, or geographic mobility, is so important that it has the power to shape landscapes. According to the 1910 US Census, more than 90% of Black US residents lived in the American South [20]. Starting in 1916, this concentration changed substantially with the start of the Black Migration; to date, the largest internal mass migration of more than six million Black southerners moving to the Northern, West, and Midwest states over 60 years [21]. Primarily due to racial discrimination and post-World War I reconstruction, this mass movement changed the landscape of urban regions across the USA; by 1970, 80% of Black residents in the USA was now living in cities [22, 23]. While migration for Black Americans leaving the South provided a social and economic improvement, it was not without a cost [24]. Smith and Welch find that even though Black Americans who migrated from the South were, on average, healthier than those who remained; they experienced higher rates of mortality later in life compared to those that stayed in the

South during this period; an effect that may still generationally precipitate within more current identified health disparities [25]. More recently, we find geographic mobility appears to be a benefit that many are not afforded. Generally, residents of disadvantaged communities are unlikely to escape the numerous adverse effects associated with high-poverty communities and are faced with almost insurmountable barriers when presented with the argument of “just move somewhere else” [26]. Longitudinal observational studies find that once a household enters into a high-poverty area, they are likely to stay or only move into similar areas elsewhere due, mainly, to limited options and the cyclic nature of poverty [27]. More specifically, these changes in the environment also resulted in more immediate changes in movement patterns within communities. During the initial 2-year 2008 recession period, over 170,000 small businesses closed across the USA [28]. These businesses are often relied on by disadvantaged neighborhood residents as sole proprietors of essential goods and services, including food and groceries. In addition, small businesses remain an important resource for employment in poor areas. With these assets removed, individuals are forced to travel elsewhere to meet needs, even if moving residents is not an option. Combined with the fact that disadvantaged communities that struggle with population-level income inequity also have residents that are more likely to experience poorer mental and physical health and face much more limited employment and educational opportunities [29–31]. Suggestions for these health disparities often revolve around inaccessibility to health centers or clinics. Positioned between preventative health and food, individuals are often forced to choose one over the other based on distance and travel-related resources.

Region is a geographic theme that most are probably familiar with. While some regions are fairly standardized and explicit with agreed-upon classifications (e.g., southern regions, mountainous regions), other regions are more contextually defined and may not be immediately obvious. For example, the Rust Belt region of the USA, called so due to the deindustrialization the region experienced starting in the mid-twentieth century, has been examined through the lens of spatial epidemiology, particularly the human–environment interaction of hazard exposure and infrastructure decline [32, 33].

Regions play a critical role in identifying spatial health inequities. Why does one geographic area experience a higher prevalence of an adverse health outcome than other similar areas? Investigating questions like this usually starts with first understanding a region’s demographics. Demographics are the aggregated individual characteristics of a population. If you consider a classic case-based study, it is typical to include subject-level characteristics such as age, gender, race/ethnicity, income, and education, factors that may influence a person’s health status. Spatial epidemiological studies are similar, but because the subjects are locations, aggregated demographics are typically included the proportion of residents under 18 years of age, the proportion of African-American or Black residents, the median household income, and the mean years of education. The inclusion of demographics in spatial epidemiological studies assists in identifying what particular populations may be experiencing health inequities. While looking at a map may tell us that the incidence of sexually transmitted infections (STIs) is highest in the southwest US region, the application of demographics would suggest an association between the region’s high STI incidence and systemic poverty [34].

**Box 14.1 What Is Spatial Health Inequity?**

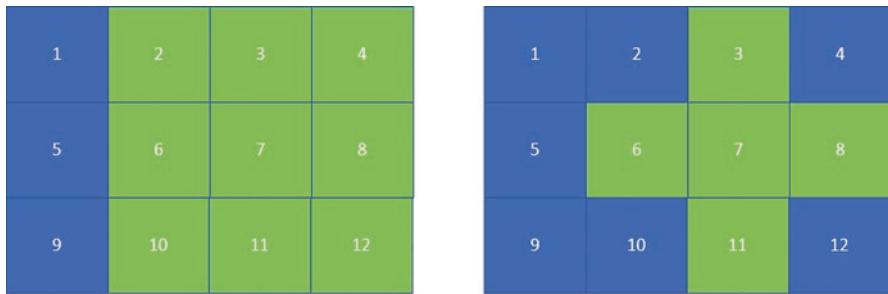
Spatial, or geographic, health inequities are systematic differences in the distribution of health outcomes experienced by people in a specific area. While spatial health inequities can be experienced by entire populations of specific areas, more commonly they are a result of the interaction between location and population demographics, i.e., an adverse health outcome experienced in a specific location by a particular proportion of the population. The application of spatial epidemiology assists in identification of spatial health inequities and allows for more effective resource allocation and directed prevention efforts.

## 2.2 *Getting to Know Your Neighbor*

As stated in the above section, what makes spatial epidemiology unique is the consideration of surrounding locations. For example, we would expect the rate of obesity to be similar among a group of surrounding neighborhoods, but if not, why not? To examine a spatial epidemiological question like this, we must first define “surrounding” neighborhoods. We must first define who our neighbors are in theory and practice.

In spatial epidemiology, neighboring locations are defined in a number of ways but are typical of the same “area unit” or spatial level, e.g., coordinates being neighbors with other coordinates or polygons being neighbors with other polygons. Among the varying ways neighbors can be defined, there are two general categories: (1) by contiguity or (2) by distance. Contiguity neighbors are locations that share one or more borders and, thus, typically are used when locations of interest are polygons. Because “sharing a border” can also be defined in a number of ways, there are varying types of contiguity, the two most popular being queen contiguity and rook contiguity, both depicted in Fig. 14.3. Named after the movements of their corresponding chess pieces, queen contiguity considers neighbors to be any location that touch borders (even corners), in contrast to rook contiguity, which only considers locations to be neighbors if they share a vertices border (no corners!).

From Fig. 14.3, you should notice that depending on if we apply queen contiguity or rook contiguity, the number of neighboring locations that box 7 has will be different, 8 or 4, respectively. Choosing which type of contiguity to apply is almost always contextual to the purpose and geography of the epidemiology interest or question. If box 7 and 2 are culturally or socially different or members of different regions, applying queen contiguity may not be the best practice. The figure above focuses only on the neighbors of box 7, which is somewhat misleading because each of the 12 squares in the figures will have a set of neighbors. For example, applying rook contiguity, the neighbors for box 1 would be 2 and 5. Applying queen contiguity, the neighbors for box 1 would be 2, 5, and 6. If we applied queen contiguity, box 1 and box 7 would not be neighbors but share a neighbor (box 6). This logic forms



**Fig. 14.3** Examples of queen and rook contiguity

the basis of a contiguity weight matrix: a list of locations, each with their own subset of neighbors. This weight matrix suggests that the prevalence of some health outcomes should be similar for box 6 and box 7 and less similar for box 7 and box 1, though because they might *share* a neighbor, they may be somewhat similar. What about the prevalence of a health outcome for box 5 compared to box 8? Regardless if we applied rook or queen contiguity, box 5 and box 8 are not neighbors and would not share any neighbors, so you would expect the prevalence to be unrelated.

Applying contiguity weights for polygon locations makes sense because a polygon has discrete borders. What if the locations are a list of coordinates? In this case, we can apply a distance weight matrix. Neighbors based on distance rather than shape are common when working with coordinated data. The same logical principles apply in a distance weight matrix as in a contiguity weight-based matrix. A distance criterion is chosen, and any two coordinates within that distance would be labeled as neighbors:

$$w_{ij}(d) = \begin{cases} 1 \\ 0 \end{cases}$$

While there exists a number of theoretical and conceptual criteria for what *distance* to choose, the simplest method is applying a minimum distance threshold, such that no location is left neighborless.

Neighborless do occur in spatial epidemiology. You may be using contiguity-based weights and have a literal island, which is also the name given to observations that have no neighbor. You may be using distance weights and have an observation that falls outside an a priori threshold. Having “islands” in spatial data can be difficult to work and require an advanced specialized methodological approach.

It should be noted that while contiguity weights are seldom, if ever, applied to coordinate locations, distance weights are sometimes used with polygon locations, which makes conceptual sense. This is typically done by applying the distance

thresholds to the calculated center of each polygon. Of course, this specific application has limitations, as it ignores the variation in polygon size, and the subsequent range of weights may bias any statistical results.

A number of more advanced weight matrices can be applied to spatial data (e.g., inverse, KNN, kernel weights) and further parameters that need to be considered among the weights discussed (e.g., Euclidean distance, great-circle, or band weights). However, the concept behind these more advanced methods is the same end product, spatial weights, where each location has a list of its neighboring locations. Further, calculating any weights by hand, while technically possible, is more efficiently done using a specialized GIS software.

### 3 Testing Spatial Dependence

In spatial epidemiology, determining the presence of autocorrelation for a health outcome is often the first step in the analysis. In Sect. 2.1, autocorrelation was briefly discussed. Here, we further explore the concept of autocorrelation and review a basic statistic test used to determine if and how spatial dependence is present.

As a reminder, autocorrelation in spatial statistics occurs when a variable is correlated with itself at different locations. That is, we expect neighboring locations to have similar values for the variable of interest. Similar to the Pearson coefficient for bivariate correlation, spatial autocorrelation is also represented with a coefficient: Moran's  $I$ . Like the Pearson coefficient, the spatial autocorrelation coefficient ranges from  $-1$  (perfect negative autocorrelation) to  $1$  (perfect positive autocorrelation), as illustrated in Fig. 14.4.

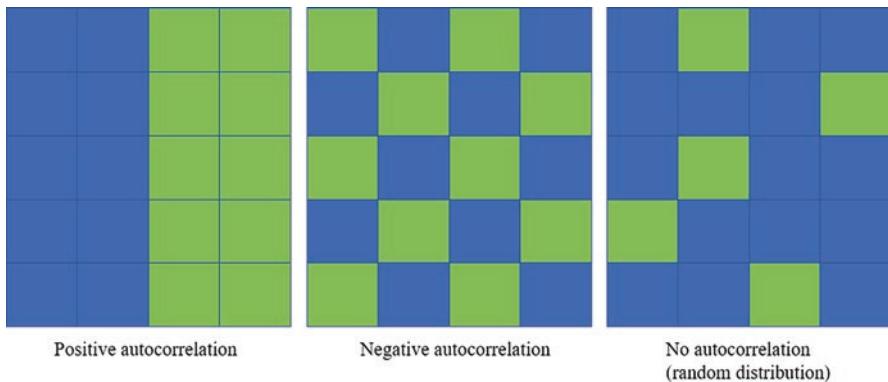
Unlike classical correlation, however, spatial autocorrelation is multidimensional and can be calculated in two distinct ways: (1) global autocorrelation and (2) local autocorrelation, which are covered below.

Global autocorrelation assumes a homogeneity throughout all the locations being examined and uses an average neighbor value within its calculation:

$$\text{Global Moran's } I = \frac{N \sum_{i=1}^N \sum_{j=1}^N w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{W \sum_{i=1}^N (x_i - \bar{x})^2}$$

In the above equation,  $N$  is the number of spatial observations,  $x$  is the variable of interest,  $w_{ij}$  is the spatial weights matrix, and  $W$  is the sum of all the weights. We can compare this to the local Moran's  $I$  coefficient calculation:

$$\text{Local Moran's } I(i) = \frac{\frac{x_i - \bar{x}}{\sum_{i=1}^N (x_i - \bar{x})^2} \sum_{j=1}^N w_{ij} (x_j - \bar{x})}{N}$$



**Fig. 14.4** Depiction of spatial autocorrelation

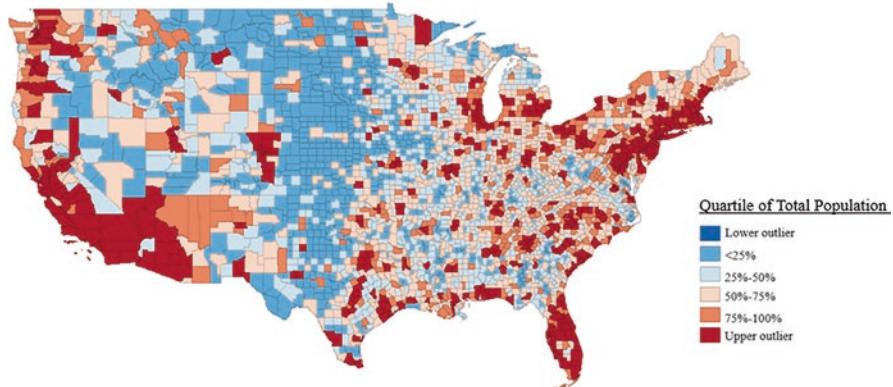
While both equations use similar variables and operations, there are also obvious differences. Solving the Global Moran's  $I$  equation will result in one single coefficient index that represents spatial autocorrelation across *all* the locations. However, due to the range of variability in the number of neighbors each location has, a single index may not be truly representative. Calculating Moran's  $I$  for each location may be more accurate with the assistance of GIS software. Rather than a single index coefficient, local autocorrelation results in a local indicator of spatial autocorrelation (LISA) map, where each location is color-coded based on membership to one of the five possible groups: (1) a location where the variable of interest is significantly higher with neighbors that also have significantly high values, (2) a location where the variable of interest is significantly lower and its neighbors also have low values, (3) a location where the variable of interest is significantly higher but is surrounded by locations with lower values, (4) a location with a significantly lower value and has neighbors with high values, or (5) locations that have values that are not significantly different than average. In this instance, the probability of significance is determined using permutation testing and the resulting  $z$  score.

## 4 Spatial Visualizations

The LISA map is a typical component of spatial epidemiology, though it is only one of many. Visualizing geographic data in spatial epidemiology is critical in disseminating and communicating results. Beyond LISA maps, the choropleth map is also one of the most widely used visualizations in spatial epidemiology, and while the name may not sound familiar, the maps likely are. An example of a choropleth map is shown in Fig. 14.5.

Choropleth maps use colors to depict the aggregation of a variable across locations. The locations in Fig. 14.5 are polygons of US counties and county equivalents. The variable of interest is the total population as defined by the Census Bureau

Total Population of U.S. Counties, 2020



**Fig. 14.5** An example of a choropleth map. (Data source: U.S. Census Bureau [35])

[35]. However, if we view the map closely, we see that the color code for each location represents the quartile of values, not the actual values. This is a specific type of choropleth map called a hinge or boxplot map that shows the quartile categories of observation. This map uses a 1.5 hinge, meaning that outliers are identified above or below 1.5 times the standard deviation. Some features in the Fig. 14.5 map should be familiar: a title, a legend, and appropriate categories for the depicted variable of interest. These features must be clear, concise, and present. Other useful features that can also be included in spatial visualizations are a compass to assist with orientation, a scale to represent physical distance, and grid lines that depict longitude and latitude.

You may be able to imagine a number of scenarios where the map in Fig. 14.5 may be helpful in spatial epidemiology. Maps like this could assist in deploying health resources based on the risk at specific locations [36]. The maps can also inform epidemiologists where infectious disease outbreaks occur and what demographics are associated [37].

## 5 Application

The information reviewed in this chapter can be combined and applied to understand spatial epidemiological studies better. In this section, we will review such a case. While spatial health data is readily available online, we will use randomly generated data and fictional locations to illustrate the concepts in this chapter better.

For this example, spatial epidemiologists are tasked with determining if residents of certain neighborhoods are more or less likely to be exposed to restaurants with poor health scores. Similar to all epidemiological exercises, the first step is inspecting the data. A preview of this fictional data is shown in Fig. 14.6. We learn from an

Restaurant ID	Coordinates	Health Score	Neighborhood
001	(1.11, -9.13)	74	A
002	(1.33, -9.12)	88	A
003	(1.35, -9.12)	90	A
004	(1.90, -9.01)	93	B
005	(1.99, -9.10)	98	C
006	(1.93, -9.09)	71	C
007	(1.93, -9.07)	93	C

**Fig. 14.6** An example of spatial data

imaginary data codebook that observations for each restaurant location among neighborhoods in an area of interest construct data. Column 1 of the data is a unique restaurant ID, column 2 of the data is (obviously fictional) coordinates for each respective restaurant, and column 3 is the most recent health score given to respective restaurants by the local health department and is based on cleanliness and required health standards meant to prevent illness due to foodborne illness, where the score ranges from 0 to 100, with 100 representing the best possible score. The last column of the data represents the restaurant's neighborhood.

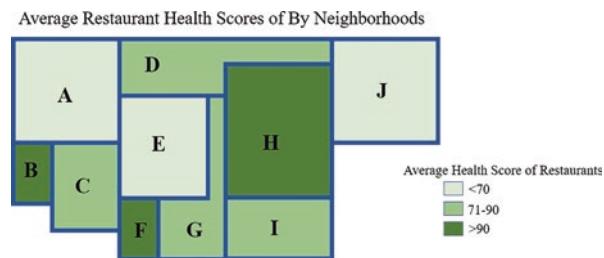
Next, we need to decide what locations are: points or polygons or what are the places of interest. Since coordinates are a named column in the data, it may be tempting to use coordinates and assume our locations will be restaurants. However, if we look closely at the objective being investigated, the focus is on neighborhoods:

### **5.1 Determining if Residents of Certain Neighborhoods Are More or Less Likely to Be Exposed to Restaurants with Poor Health Scores**

If the objective was *determining if spatial dependence exists among restaurants according to health scores*, then using points of restaurant locations may be more appropriate. However, since our objective is at the neighborhood level, we must proceed with some data integration.

Combining or integrating epidemiological data with a spatial data set is common in spatial epidemiology and usually starts with utilizing a shapefile using GIS software. A shapefile is an electronic file containing data and metadata related to a collection of specific geographic locations; points or polygons could contain information on one location or thousands. For example, a shapefile may contain spatial vector data of all the census tracts in Johnson County, Iowa, it may contain all census tracts in the USA, it may contain all nations of the world, or it may

**Fig. 14.7** Choropleth map of the average restaurant in each neighborhood



contain all fitness centers in Los Angeles, California. For our example, we will use a shapefile containing a neighborhood collection. We can use a shapefile of our neighborhoods and the data from Fig. 14.6 to calculate the average health score of restaurants within each neighborhood. We can then merge the aggregated health score data with our shapefile of neighborhoods to make a choropleth map, as depicted in Fig. 14.7.

You may notice, much like real locations, our imaginary neighborhoods vary in size and thus likely vary in the number of restaurants and population, making our average health score calculation ill-fitting and biased. To account for this, many GIS software allows for more sophisticated aggregations such as standardized rate calculations where population and area can be accounted for. For our map in Fig. 14.7, we can use it to make preliminary inferences. Neighborhoods B, F, and H appear to have restaurants with the best health scores, and neighborhoods A, E, and J appear to have restaurants with the lowest average health scores. While this simple aggregation and visualization may be preliminarily helpful, we can continue examining whether these scores are associated with neighborhoods.

For our next step, we need to apply a neighbor weight matrix and decide that queen contiguity is the most appropriate. Thus, our weights would look like: A [B, C, D, E, F], B [A, C], C [A, B, E, F], D [A, E, G, H, J], E [D, A, C, F, G, I, H], F [E, C, G], G [D, E, F, I, H], H [D, G, I, J], I [H, G], and J [D, H]. If you remember from the previous section, this weight matrix would imply that we expect, for example, neighborhood F to have similar characteristics as it is neighbors, the neighborhoods of E, C, and G. We can use these waits then to determine the global autocorrelation of average restaurant health scores among neighborhoods.

Let us imagine that we use our GIS application and determine the global Moran's  $I$  of average health score per neighborhood to be 0.371, and permutation testing gives us a  $Z$  value of 36.0 ( $p < 0.001$ ). We could interpret this to mean that, overall, the average restaurant health score of restaurants per neighborhood is spatially dependent. Because the index is positive, locations are likely to have similar scores. For example, the average health score in neighborhood G should be similar among its neighbors. In the context of spatial epidemiology, this information may provide

helpful health programs and prevention efforts, e.g., if a neighborhood has lower restaurant health scores, the surrounding neighborhoods are significantly likely to have similar lower scores. They may necessitate broadening efforts to prevent food-borne illnesses or the deployment of restaurant staff training.

A spatial epidemiological investigation, like our example, may not stop here and deploy further advanced methods. Epidemiologists may re-examine results using a different neighbor matrix method based on context and purpose. In addition, there are also a number of further, more complex steps that can be taken in the analysis, such as local autocorrelation testing or rate change analysis. Still, the above example gives a basic-level description of a typical spatial epidemiological application.

## 6 Further Practice

1. The presence of spatial dependence in variables among locations can be determined using a choropleth map.
  - (a) True
  - (b) False
2. When would a group of locations be considered a region?
  - (a) If they had similar geographic or population characteristics
  - (b) If they were contiguous
  - (c) If they were all within the same administrative boundary
  - (d) Both (a) and (b)
  - (e) None of the above
3. What is the primary difference between location and place?
4. Which of the below is an example of a place?
  - (a) A grocery store
  - (b) The city of Atlanta, GA
  - (c) The Mississippi River
  - (d) All of the above
  - (e) None of the above
5. Explain why population movement is relevant to spatial epidemiology?
6. Using the image below and applying a rook contiguity would result in box 12 having what neighbors?
  - (a) 7, 8, and 11
  - (b) 8 and 11
  - (c) 9, 10, and 11
  - (d) Cannot be determined

1	2	3	4
5	6	7	8
9	10	11	12

7. Which is the most appropriate description of spatial health inequity?
- (a) An adverse health outcome due to toxic environmental exposure
  - (b) Differences in prevalence or incidence of a health outcome associated with location
  - (c) Differences in prevalence or incidence of a health outcome due to location
  - (d) Differences in health outcomes among populations in the same location
  - (e) None of the above
8. Using contiguity method of neighbors, the relationship between neighboring locations can be described as which of the following?
- (a) Binary (either they are neighbors or they are not)
  - (b) Continuous (they can be proportionally neighbors)
  - (c) Categorical (different levels of neighbors)
  - (d) None of the above
9. A location with no neighbors is typically called what?
- (a) A loner
  - (b) A stranger
  - (c) An island or isolate
  - (d) Nothing, every location has a neighbor
  - (e) None of the above
10. Among a group of point coordinates, the largest distance between any two points is 100 km, the shortest distance between any two points is 25 km. What distance band would be applied for a distance weight matrix so that no point was left neighborless?
- (a) 100 km
  - (b) 25 km
  - (c) 62.5 km
  - (d) Depends on the number of total points

11. Which of the following statements best describes a positive global Moran's coefficient.
  - (a) As one variable increases, a different variable also increases.
  - (b) If one variable gets larger, the same variable gets larger in neighboring locations.
  - (c) The value of a variable at one location will have a similar value in neighboring locations.
  - (d) If one variable gets smaller, the same variable gets larger in neighboring locations.
  - (e) All of the above are false.
12. A local spatial autocorrelation test is performed only if a global autocorrelation is significant.
  - (a) True
  - (b) False
13. A global Moran's  $I$  value of  $-0.32$  would indicate what?
  - (a) The value of a variable at a specific location is likely to be dissimilar among neighboring locations.
  - (b) The value of a variable at a specific location is significantly lower than the value among neighboring locations.
  - (c) The value of one variable will be dissimilar compared to the value of another variable at the same location.
  - (d) There is an error in the results.
14. The presence of spatial dependence may change based on which neighboring method is used.
  - (a) True
  - (b) False
15. Which of the following features should always be included in map visualizations?
  - (a) A title
  - (b) A legend
  - (c) A compass
  - (d) Both (a) and (b)
  - (e) Both (b) and (c)

### Answer Keys

1. (b)
2. (d)
3. A location identifies a geographic space. A place is the description of what occupies a location
4. (d)

5. Populations are not static and interact with multiple environments in multiple locations. Understanding these interactions gives insight into health outcomes
6. (b)
7. (b)
8. (a)
9. (c)
10. (a)
11. (c)
12. (b)
13. (a)
14. (a)
15. (d)

## References

1. Vicentini CB, Contini C. Control measures of a 400-year-old plague epidemic: an example of past efficiency at controlling disease and similarities with current epidemics. *Infez Med.* 2020;28(4):621–33.
2. Snow J. On the mode of communication of cholera. London: John Churchill; 1855.
3. Lapp AD. Treating tuberculosis in the dry climate belt. *Can Med Assoc J.* 1925;15(8):819–22.
4. Mark S. Early human migrations (ca. 13,000 years ago) or postcontact Europeans for the earliest spread of *Mycobacterium leprae* and *Mycobacterium lepromatosis* to the Americas. *Interdiscip Perspect Infect Dis.* 2017;2017:6491606.
5. Scroggins S, Shacham E. Using spatial-peak analysis to model COVID-19 outbreaks. *Ann Epidemiol.* 2022;75:73.
6. Tobler W. A computer movie simulating urban growth in the Detroit region. *Econ Geogr.* 1970;46:234–40.
7. Feurer T, Sauerbrey R. 8 – Characterization of short laser pulses. In: Dunning FB, Hulet RG, editors. Experimental methods in the physical sciences, vol. 29. San Diego: Academic Press; 1997. p. 193–229.
8. Barbusjani G. Autocorrelation of gene frequencies under isolation by distance. *Genetics.* 1987;117(4):777–82.
9. Tomczuk R, Miller DM. Near-optimal PLA input variable pairing using autocorrelation techniques. *Microelectron J.* 1992;23(7):523–31.
10. Hogenraad R, McKenzie DP, Martindale C. The enemy within: autocorrelation bias in content analysis of narratives. *Comput Hum.* 1996;30(6):433–9.
11. Association of American Geographers. Guidelines for geographic education. Elementary and secondary schools. Washington, DC: National Council for Geographic Education; 1984.
12. Kelly JD, Park J, Harrigan RJ, Hoff NA, Lee SD, Wannier R, et al. Real-time predictions of the 2018–2019 Ebola virus disease outbreak in the Democratic Republic of the Congo using Hawkes point process models. *Epidemics.* 2019;28:100354.
13. Charu V, Zeger S, Gog J, Bjørnstad ON, Kissler S, Simonsen L, et al. Human mobility and the spatial transmission of influenza in the United States. *PLoS Comput Biol.* 2017;13(2):e1005382.
14. Humpel N, Owen N, Leslie E. Environmental factors associated with adults' participation in physical activity: a review. *Am J Prev Med.* 2002;22(3):188–99.
15. Lewis NO, Lapham SC, Skipper BJ. Drive-up liquor windows and convicted drunk drivers: a comparative analysis of place of purchase. *Accid Anal Prev.* 1998;30(6):763–72.
16. Harbers MC, Beulens JWJ, Boer JMA, Karssenberg D, Mackenbach JD, Rutters F, et al. Residential exposure to fast-food restaurants and its association with diet quality, overweight

- and obesity in the Netherlands: a cross-sectional analysis in the EPIC-NL cohort. *Nutr J.* 2021;20(1):56.
- 17. Bartholy J, Pongrácz R. A brief review of health-related issues occurring in urban areas related to global warming of 1.5°C. *Curr Opin Environ Sustain.* 2018;30:123–32.
  - 18. Santamouris M, Cartalis C, Synnefa A, Kolokotsa D. On the impact of urban heat island and global warming on the power demand and electricity consumption of buildings—a review. *Energ Buildings.* 2015;98:119–24.
  - 19. Lynch DH, MacRae R, Martin RC. The carbon and global warming potential impacts of organic farming: does it have a significant role in an energy constrained world? *Sustainability.* 2011;3(2):322–62.
  - 20. Gibson C, Jung K. Historical census statistics on population totals by race 1790 to 1900, and by Hispanic origin, 1970 to 1900 for the United States, regions, divisions, and states. Washington, DC: United States Census Bureau; 2002.
  - 21. Lemann N. The promised land: the great black migration and how it changed America. New York: Alfred A. Knopf; 1991.
  - 22. Wilkerson I. The long-lasting legacy of the great migration. Smithsonian; 2016.
  - 23. Trotter JW, editor. Great migration: an interpretation. New York: Oxford University Press; 2005.
  - 24. Smith JP, Welch F. Black economic progress after Myrdal. *J Econ Lit.* 1989;27(2):519–64.
  - 25. Black D, Sanders S, Taylor L. The impact of the Great Migration on mortality of African Americans: evidence from the Deep South. *Am Econ Rev.* 2015;105(2):477–503.
  - 26. Sampson R. Individual and community economic mobility in the great recession era. Paper originally presented at the Federal Reserve conference on Economic Mobility. Washington, DC: Harvard University; 2015.
  - 27. Comey J, Souza X, Weismann G. Struggling to stay out of high-poverty neighborhoods: lessons from the moving to opportunity experiment. Report no.: 6. Washington, DC: The Urban Institute; 2008.
  - 28. Sahin A, Kitao S, Cororaton A, Laiu S. Why small businesses were hit harder by the recent recession. *Econ Finan.* 2010;17(4):1–7.
  - 29. Leventhal T, Brooks-Gunn J. Moving to opportunity: an experimental study of neighborhood effects on mental health. *Am J Public Health.* 2003;93(9):1576–82.
  - 30. Hamm L, McDonald S. Helping hands: race, neighborhood context, and reluctance in providing job-finding assistance. *Sociol Q.* 2015;56:539–57.
  - 31. Gaskin DJ, Thorpe RJ Jr, McGinty EE, Bower K, Rohde C, Young JH, et al. Disparities in diabetes: the nexus of race, poverty, and place. *Am J Public Health.* 2014;104(11):2147–55.
  - 32. Rajaei M, Echeverri B, Zuchowicz Z, Wiltfang K, Lucarelli JF. Socioeconomic and racial disparities of sidewalk quality in a traditional rust belt city. *SSM Popul Health.* 2021;16:100975.
  - 33. Lynch EE, Meier HCS. The intersectional effect of poverty, home ownership, and racial/ethnic composition on mean childhood blood lead levels in Milwaukee County neighborhoods. *PLoS One.* 2020;15(6):e0234995.
  - 34. Harling G, Subramanian S, Bärnighausen T, Kawachi I. Socioeconomic disparities in sexually transmitted infections among young adults in the United States: examining the interaction between income and race/ethnicity. *Sex Transm Dis.* 2013;40(7):575–81.
  - 35. U.S. Census Bureau. 2015–2020 American community survey 5-year public use microdata samples. 2021. Available from: <https://data.census.gov/>.
  - 36. Scroggins S, Goodson J, Afrose T, Shacham E. Spatial optimization to improve COVID-19 vaccine allocation. *Vaccine.* 2023;11(1):64.
  - 37. Shacham E, Nelson EJ, Schulte L, Bloomfield M, Murphy R. Condom deserts: geographical disparities in condom availability and their relationship with rates of sexually transmitted infections. *Sex Transm Infect.* 2016;92(3):194.

# Chapter 15

## Survival Analysis and Applications Using SAS and SPSS



Rafiqul Chowdhury and Shahariar Huda

### 1 Introduction

This chapter presents a brief overview to survival analysis and some data analysis approaches using specific epidemiologic and other data. Also, these approaches include problems adopted by survival analysis, including the outcome variable, and the consideration of “censored data” or “failure time data.” In addition, a few concepts are explained, such as the survival function, hazard function, data preparation for survival analysis, and the goals of survival analysis. Finally, real-life data examples are considered to demonstrate survival analysis methods using SAS and SPSS.

### 2 The Time-to-Event Data

What analytic problems are addressed by survival analysis? Survival analysis is a statistical method which is used for analyzing the response variable in the time until an event occurs. The *event* is a variable of interest which can be death, disease occurrence, relapse from remission, recovery, or any defined experience of interest that may occur in an individual within a given time, whereas *time* can be duration of years, months, weeks, or days from the beginning of the study of an individual till the occurrence of an event. Also, an individual’s age when an event occurs can be used as the time variable.

---

R. Chowdhury ()

Shannon School of Business, Cape Breton University, Sydney, NS, Canada  
e-mail: [Rafiqul\\_chowdhury@cbu.ca](mailto:Rafiqul_chowdhury@cbu.ca)

S. Huda

Kuwait University, Kuwait City, Kuwait  
e-mail: [shahariar.shamsulhuda@ku.edu.kw](mailto:shahariar.shamsulhuda@ku.edu.kw)

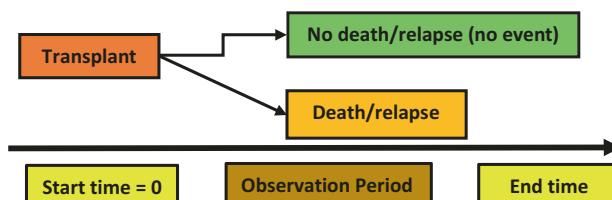
This time variable is generally called survival time because the individual has survived without any event that occurred over some follow-up period. However, it is also typical to represent the *event* as a *failure* and the *time* as the *failure time*. Hence, survival analysis is the *failure time* data analysis. It is possible to consider more than one event in the same analysis. However, we will consider that only one event can occur in a given time in an individual.

Although it is possible to consider more than one event (e.g., more than one antenatal complication), the statistical problem is either a recurrent event or a competing risk problem, which is beyond the scope of this chapter. As mentioned earlier, typically, these types of data are known as *survival data* or *failure time data*. Incidentally, failure time data are also of concern to engineers, and the analysis of such data is called “*Reliability Analysis*” in the case of scientists in the engineering field.

## 2.1 A Real-Life Example

Consider a data set of 2204 patients from the European Group for Blood and Marrow Transplantation (EBMT) study. These patients received bone marrow transplantation between 1995 and 1998. The event of interest here is death corresponding to both relapse and death after the bone marrow transplantation (Fig. 15.1). This data set is available from the “mstate” R package [1]. Out of 2204 patients, 458 had experienced the event of interest, that is, relapsed or died.

As a second example, consider the data on maternal morbidity in Bangladesh by Akhter et al. [2]. The survey was conducted from November 1992 to December 1993 by the Bangladesh Institute for Research for Promotion of Essential and Reproductive Health Technologies (BIRPERHT). The study subjects comprised pregnant women with less than 6 months of pregnancy. All the selected pregnant women were followed regularly (roughly at an interval of 1 month) throughout the pregnancy. In addition, data were collected on antenatal complications (e.g., hemorrhage, edema, excessive vomiting, fits/convulsion, and delivery complications), along with other sociodemographic variables. The study sample comprised 993 pregnant women. Among them, 485 women were free from antenatal complications, while 508 women suffered antenatal complications during the pregnancy.



**Fig. 15.1** Relapse and death after the bone marrow transplantation

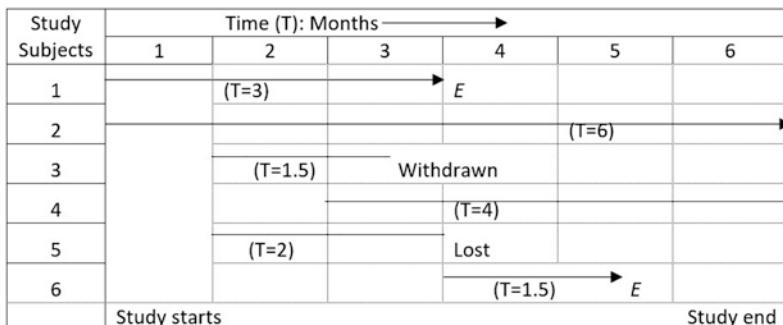
### 3 Censored Data

A typical feature of survival data is that we may not be able to observe the occurrence of the event (e.g., death) for all the patients by the end of the observation period. Therefore, times required for some patients' actual event occurrence (or survival times) are unknown. In other words, some of the data are censored. For example, in the previous EBMT data, only 458 patients had experienced the event of interest during the study period. Hence, the exact time of the event is known for those 458 patients. However, for the remaining patients, we only recorded the time till the end of the observation period (up to which they were event free). Therefore, survival times for the event-free patients are censored, that is, we do not know the exact survival times (event times) of some people. Most survival analysis methods must consider the problem of censoring in the analysis to allow for valid inferences.

Furthermore, the distribution of survival times is usually skewed, limiting analysis methods that require the assumption of normally distributed data. For the analysis of survival (or time-to-event) data, it is necessary to consider the occurrence of the event (outcome) of interest, as well as when (the time) that event occurred. For example, statistical methods such as logistic and linear regression cannot simultaneously consider survival data's *event* and *time* aspects. Traditional regression methods are also not equipped to handle censoring.

There are generally three reasons for censoring: (i) a subject may not experience the event during the study period; (ii) a subject is lost to follow-up during the study period; and (iii) a person withdraws from the study because of death (if death is not the event) or some other reason (e.g., adverse drug reaction or other competing risks). Figure 15.2 explains the different types of censoring experienced by the six study subjects in a hypothetical study. An *E* denotes the event of interest observed by a subject, and *T* indicates the time in months.

Figure 15.2 illustrates the experience of six subjects observed over 6 months (time). Again, the notation *E* is used to identify that an individual experienced an event.



**Fig. 15.2** Different types of censoring experienced by the six study subjects

In Fig. 15.2, subject 1 is observed from the start of the study till the event occurrence at month 3, with a survival time of 3 months; hence not censored.

Subject 2 is observed from the beginning to the end of the sixth month without the event's occurrence, and the survival time is censored because we can only say that the time is at least 6 months.

Subject 3 entered the study at the beginning of the second month and withdrew from the study at 3.5 months. Therefore, this person's survival time is 1.5 months and is censored.

The fourth subject enters at month 3 and remains event free till the end of the study. This person's censored time is 4 months.

Subject 5 enters the study at month 2 and is followed until month 3, then he is lost to follow-up. The censoring time for this subject is 2 months.

Subject 6 enters at month 4 and is followed until the event at month 5.5. As with subject 1, this subject also has an event; the survival time is 1.5 months.

Table 15.1 presents the data and layout for the six subjects from the hypothetical study shown in Fig. 15.2 for survival analysis. For each subject, the survival time and censorship status are shown in the table. In Table 15.1, the “Censoring” column indicates whether this time is censored (1 = an event and 0 = censored).

For example, subject 1 has a survival time of 3.0 and a censoring indicator of 1, whereas for subject 3, the survival time is 1.5 and the censorship indicator is 0.

### 3.1 Types of Censoring

In survival analysis data, generally three different types of censoring are encountered. These are left-censoring, right-censoring, and interval-censoring. Among these, the third type is most common in real life.

When during the follow-up period, the survival time of a subject remains incomplete on the upper end of the period, the censoring is called right-censoring. For example, for each of the four subjects in Fig. 15.2, the survival time is incomplete on the right side. The censoring happens because the study ends before the failure (death) or the subject is lost to follow-up having been withdrawn from the study before it ends. The actual (unobserved) survival time of these individuals may extend beyond the upper end of the interval.

**Table 15.1** Data and layout for the six subjects

Subject	Time	Censoring
1	3.0	1
2	6.0	0
3	1.5	0
4	4.0	0
5	2.0	0
6	1.5	1

Left-censoring occurs when an individual's true survival time is equal to or less than the survival time observed/recoded. For example, in a study of COVID-19 infection among the elderly, a failure may be recorded when a subject tests positive (for the virus) for the first time. However, the actual time of exposure to the virus may not be known and hence the actual failure time may also remain unknown and may be less than or equal to the recorded failure time. Hence, there is left-censoring of the survival time. In other words, when left-censoring of a person's survival time happens at a time  $t$ , the implication is that the exact time of failure is unknown but the failure occurred before time  $t$ .

When a subject's failure occurs within a specified interval of time, but the actual time of occurrence is unknown, it is said that interval-censoring has taken place. Consider the COVID-19 example. Suppose a subject undergoes the test twice, the first test at time  $t_1$  when the result is negative and then the second test at time  $t_2$  when the result is positive. Thus, the subject's failure occurred at an unknown time in the interval  $(t_1, t_2)$  and the survival time is interval-censored, with  $(t_1, t_2)$  being the relevant time interval.

In the survival analysis, one needs to consider three assumptions regarding censoring: (i) independent censoring, (ii) random censoring, and (iii) non-informative censoring. More details can be found elsewhere [3–5].

## 4 Terminology and Notation

Let  $T$  be the random variable subject's survival time, and  $t$  denotes the specific value of the random variable  $T$ . Then, the possible values of  $T$  include all non-negative numbers ( $T \geq 0$ ) since  $T$  represents time. Next, define the random variable  $d$  with the two possible values (0 or 1), indicating either the event of interest occurred (1) or censored (0). Also, define two other quantities considered in any survival analysis: the survivor function, denoted by  $S(t)$ , and the hazard function, denoted by  $h(t)$ . The survivor function  $S(t)$  gives the probability that a subject survives longer than some specified time  $t$ ,  $S(t) = P(T > t)$ .

### 4.1 The Survival Function

For any continuous random variable (r.v.)  $T$  with a probability density function (pdf)  $f(t)$ , the cumulative distribution function (cdf)  $F(t)$  is defined as

$F(t) = P(T \leq t) = \int_{-\infty}^t f(x)dx$  and the survival (survivor) function  $S(t)$  is the complement of cdf, that is,

$$S(t) = 1 - F(t) = 1 - P(T < t) = P(T > t) = \int_t^{\infty} f(x)dx.$$

It follows that  $f(t) = -\frac{dS}{dt}$ . The pdf  $f(t)$  is non-negative, and  $f(t)dt$  is approximately the probability that  $T$  will fall in the interval  $(t, t + dt)$ , that is,  $f(t)dt \approx P(t < T < t + dt)$ .

In the survival analysis,  $T$  represents the survival time and is necessarily a non-negative continuous r.v. Then, the total area under the curve of  $f(t)$  between 0 and  $\infty$  is one, that is,  $\int_0^\infty f(t)dt = 1$ .

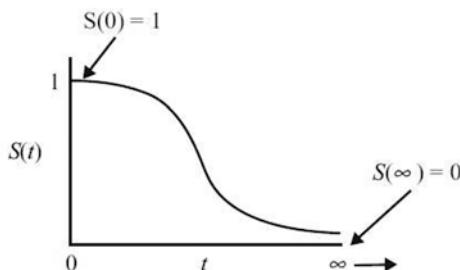
For analyzing survival data, it is fundamental to explore the survival function which provides the probability of survival at different time points. In principle, a smooth curve can be graphed representing the survival function  $S(t)$  over the range  $(0 < t < \infty)$ . The following are important characteristics of the survival function:

- (i) It is nonincreasing, that is, as  $t$  increases,  $S(t)$  has a downward trend.
- (ii)  $S(t) = 1$  at  $t = 0$ , that is,  $S(0) = 1$  since no event (failure/death) is experienced at the onset of the study and the probability of survival beyond time  $t = 0$  is unity.
- (iii)  $S(\infty) = 0$ , that is, at time  $t = \infty$ , the probability of survival is zero since if the study period is in principle extended indefinitely, none of the subjects would be alive by end of the period.

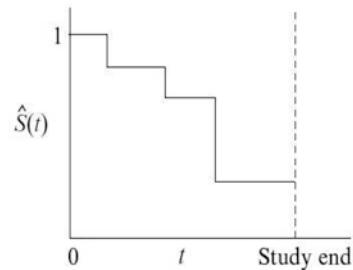
In practice, a survival function is constructed using data and the empirically constructed survival function is necessarily a step function which is nonincreasing. Furthermore, a study period can never be extended indefinitely, and there may be “competing risks.” Also, the “event” of interest may not be experienced by all the subjects in a study. Thus, the estimated survival function is only an approximation to the true survival function.

Figure 15.3 displays the smooth (theoretical) and step function (from data) survival curves. In Fig. 15.3, the hat on the survival function represents that it is estimated from data.

Theoretical  $S(t)$ :



$\hat{S}(t)$  in practice:



**Fig. 15.3** Two survival curves

## 4.2 Hazard Function

The hazard function, denoted by  $h(t)$ , is given by the following formula:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t}.$$

If  $T$  is a continuous random variable, then,

$$h(t) = \frac{f(t)}{S(t)} = -d \ln[S(t)] / dt.$$

The cumulative hazard function  $H(t)$  is another quantity, defined by

$$H(t) = \int_0^t h(u) du = -\ln[S(t)].$$

The hazard function provides, in other words, the instantaneous probability of an event occurrence at time  $t$ , per unit of time, given that the individual has survived up to time  $t$ .

While survival is the focus of survival function, the hazard function emphasizes on failure (death/event). The hazard function is also called the “conditional failure rate” since the formula gives the probability that the survival time  $T$  will fall within the small interval  $(t, t + \Delta t)$  given that  $T \geq t$ . The hazard function, also called hazard rate, has different names in different contexts. For example, in demography, it is called the “force of mortality”; in epidemiology, it is known as “age-specific failure rate”; in stochastic processes, it is called the “intensity function”; and in economics, it is referred to as the “inverse of Mill’s ratio.”

In the formula for hazard function, the expression on the right of the limit sign has the conditional probability as numerator, while  $\Delta t$  in the denominator is a small interval of time. The ratio is therefore no longer a probability but a rate, namely, “probability per unit time.” It is not confined to the range  $[0, 1]$  but can vary over the entire non-negative part of the real line. It can vary from 0 to  $\infty$  and depends on the unit of measurement for time. For any given value of  $t$ , the hazard rate  $h(t)$  satisfies  $0 \leq h(t) < \infty$ .

Many investigators find the survival function  $S(t)$  very useful because it can be used to describe directly the survival experience of a cohort of subjects under study. The hazard function  $h(t)$  also has useful applications. In contrast to  $S(t)$  being a cumulative measure,  $h(t)$  is a “measure of instantaneous potential.” The hazard rate can be very useful in identifying specific probability distribution (e.g., exponential, gamma, and Weibull) suitable as a model for the data in hand. Statistical models for survival data are often expressed in terms of the hazard function.

### 4.3 Relationship Between $h(t)$ and $S(t)$

There exists a clear direct mathematical relationship between the hazard function  $h(t)$  and the survival function  $S(t)$ . If the hazard function  $h(t)$  is known, one can derive the survival function  $S(t)$  and vice versa. For example, a constant hazard function  $\lambda$  is the characterizing property of the exponential distribution having the survival function  $S(t) = e^{-\lambda t}$ . More generally, one can express the relationship between  $S(t)$  and  $h(t)$  using either of the following formulae:

$$S(t) = \exp \left[ - \int_0^t h(u) du \right] \quad \text{and} \quad h(t) = - \left[ \frac{dS(t)/dt}{S(t)} \right].$$

### 4.4 Kaplan–Meier Estimator

We use the Kaplan and Meier (1958) estimator (also called the product-limit estimator) as the standard for survival function (survival probability) estimation [6]. Suppose we observe  $n$  subjects for some time for the occurrence of some event of interest. So, we will have a time-to-event random variable and an indicator variable denoting whether the time censored or an event. In the survival data, multiple events can occur at the same time point, which is called *ties*. To allow for possible ties in the data, suppose that the events occur at  $D$  different times  $t_1 < t_2 < \dots < t_D$  and that at time  $t_i$ , there are  $d_i$  events. Let  $Y_i$  be the number of individuals at risk at the time  $t_i$ . That is,  $Y_i$  is the number of individuals alive at  $t_i$  or who experienced the event of interest at  $t_i$ . We assume that the potential censoring time is unrelated to the possible event time. The quantity  $d_i/Y_i$  estimates the conditional probability that an individual who survives just before time  $t_i$  experiences the event at time  $t_i$ . Kaplan–Meier survival estimation, the log-rank test, and the Cox model rely on an assumption of independent censoring for valid inference in the presence of right-censored data. The Kaplan–Meier estimator is defined as follows for all values of  $t$  in the range where there is data [6]:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1, \\ \prod_{t_i \leq t} \left[ 1 - \frac{d_i}{Y_i} \right], & \text{if } t_1 \leq t. \end{cases}$$

This estimator is not well defined for values of  $t$  beyond the largest observation time. The product-limit estimator is a step function with jumps at the observed event times. The size of these jumps depends on the number of events observed at each event time  $t_i$  and the pattern of the censored observations before  $t_i$ . Using Greenwood's formula, one can estimate the variance of the product-limit estimator:

$$\hat{V}[\hat{S}(t)] = \hat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}.$$

The standard error of the product-limit estimator is  $\{\hat{V}[\hat{S}(t)]\}^{\frac{1}{2}}$ . We will illustrate

the Kaplan–Meier (product limit) method using the data set on maternal morbidity example described earlier. For illustration, we used a part of the data set (a random sample of 20%, 188 women). We recorded the time to the first incident of hemorrhage (hemodur—time-variable). Here, the occurrence of hemorrhage is our event of interest (1 = yes or 0 = no), “yes” denotes the event has occurred, and “no” represents the time variable is censored. Data for some of the selected subjects are shown in Table 15.2. Out of 188 women, 31 experienced the event and the remaining 157 were without any event, producing 83.51% censored observations. The SAS output of product-limit survival estimates is presented in Fig. 15.4 (partially). SAS codes used are provided in Sect. 4.

In Fig. 15.4, data provided in each row of the table presents time—the point of time starts from the time in the “Hemodur” column for that row and finishes just before the time in the “Hemodur” column in the next row having a different “Hemodur” value. For example, the time interval represented by the first row starts from 0 months and finishes just before 2 months (of the next row). As we can find that during this interval, there are 188 women to begin with who were at risk and that no one had the event of hemorrhage, as “Observed Events” = 0; the estimate of the “Survival” function for this time is 1.0000.

**Table 15.2** Data for some selected subjects

id	Hemodur	Hemorrhage
155	8	No
161	4	Yes
164	8	No
181	8	No
197	8	No
202	8	No
233	4	Yes
234	6	Yes
248	8	No
249	6	Yes
254	8	No
273	8	No
286	8	No
289	8	No
293	2	Yes
299	8	No
301	8	No
303	4	Yes
305	6	No

The SAS System								
The LIFETEST Procedure								
Product-Limit Survival Estimates								
Hemodur	Number at Risk	Observed Events	Survival	Failure	Survival Standard Error	Number Failed	Number Left	
0.0000	188	0	1.0000	0	0	0	188	
2.0000	.	.	.	.	.	1	187	
2.0000	.	.	.	.	.	2	186	
2.0000	188	3	0.9840	0.0160	0.00914	3	185	
2.0000 *	.	0	.	.	.	3	184	
4.0000	.	.	.	.	.	4	183	
4.0000	.	.	.	.	.	5	182	
4.0000	.	.	.	.	.	6	181	
4.0000	.	.	.	.	.	7	180	
4.0000	.	.	.	.	.	8	179	
4.0000	.	.	.	.	.	9	178	
4.0000	.	.	.	.	.	10	177	
4.0000	.	.	.	.	.	11	176	
4.0000	.	.	.	.	.	12	175	
4.0000	.	.	.	.	.	13	174	
4.0000	.	.	.	.	.	14	173	
4.0000	184	12	0.9199	0.0801	0.0198	15	172	
4.0000 *	.	0	.	.	.	15	171	
4.0000 *	.	0	.	.	.	15	170	
4.0000 *	.	0	.	.	.	15	169	
4.0000 *	.	0	.	.	.	15	168	

Fig. 15.4 Kaplan–Meier estimate

During the next interval, starting from 1 month to just before 4 months, three pregnant women experienced the event, as indicated by the “Number Failed” column in the fourth row. The **Survival** column presents probabilities which are *unconditional*—it is the probability of surviving from the beginning of follow-up time upto the time (number of months) shown in the Hemodur column.

From “Hemodur” = 2 in the fifth row, you can see that no one had the event as the column “Observed Ents” = 0. This is a censored observation, further indicated by a “\*” (star) sign in the second column. The number of subjects censored

contribute to the survival function until they drop out of the study, without being counted as a failure. In the forth row, you can find that the *unconditional* probability of surviving (under the column “Survival”) beyond 2 months = 0.9840, and the hazard rate (Failure) = 0.0160.

Figure 15.5 shows the Kaplan–Meier estimate of the survival function showing how the survival function changes over time. The step function drops when an event occurs at a particular time, whereas the graph continues flat between the time of failure. In this chart, the survival function drops steeply at the beginning of the fourth and sixth months. The vertical ticks on the chart represent censored observations. The survival probability does not change for a censored observation. For the longest follow-up period of censored observation, the survival function does not reach 0 and remains at the survival probability that is estimated at the previous interval.

The **product-limit estimator** is a classic method of estimating the survival function for right-censored data. One can also estimate the cumulative hazard function by using the following equation,  $H(t) = -\ln [S(t)]$ . Here, the estimator  $\hat{H}(t) = -\ln [\hat{S}(t)]$ . Nelson (1972) suggested an alternate method of estimating the cumulative hazard rate [7]. As Nelson mentioned, “the method is efficient for data having small sample sizes, when compared with the Product-Limit estimator.” Aalen (1978) derived the estimator of Nelson using modern counting techniques [8]. Hence, the estimator is called the Nelson–Aalen estimator of cumulative

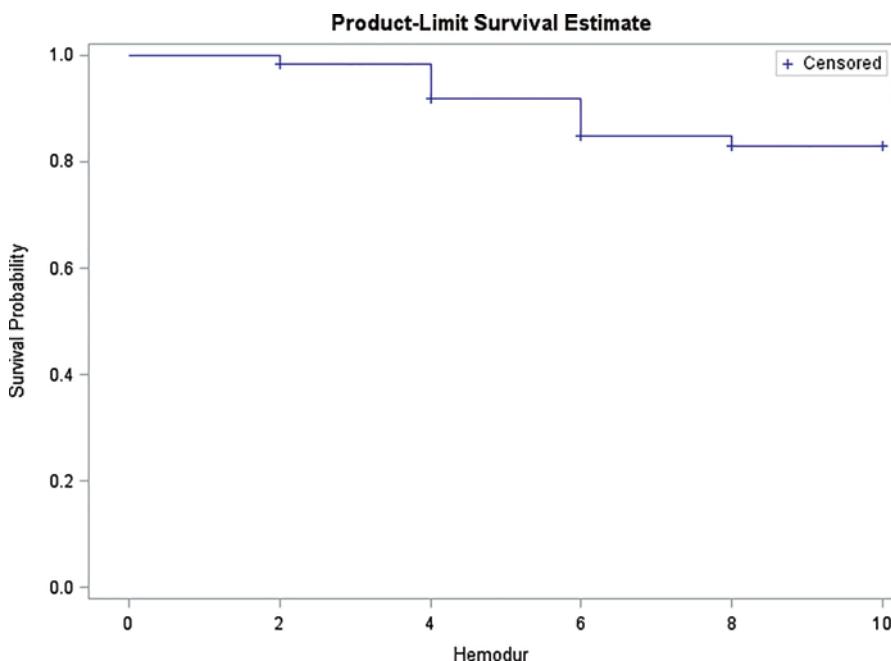


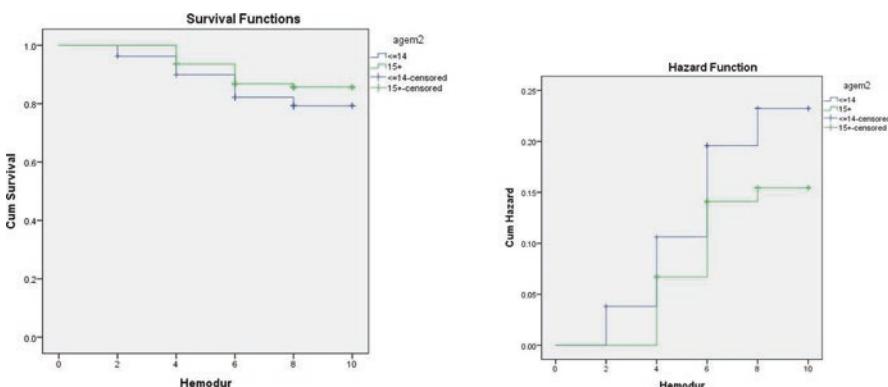
Fig. 15.5 Kaplan–Meier estimate of the survival function curve

hazard (in short, Nelson–Aalen estimator) [9]. **Nelson–Aalen estimator** is a non-parametric test which is used for estimating cumulative hazard rate function from censored data. Based on the Nelson–Aalen estimator of the cumulative hazard rate, an estimator of the survival function is given by the following equation,  $\hat{S}(t) = \exp[-\hat{H}(t)]$ .

The Nelson–Aalen estimator has two principal uses in analyzing data. First, it is used in selecting between parametric models for the time to event. Second, it provides crude estimates of the hazard rate  $h(t)$ .

## 5 Log-Rank Test

In survival data analysis, we may need to evaluate whether or not Kaplan–Meier curves for two or more groups are statistically equivalent. The most popular testing method is called the **log-rank test**. Log-rank test is, in fact, a Chi-square test, which provides an overall comparison of the Kaplan–Meier curves. Similar to a Chi-square test, the log-rank test uses observed versus expected cell counts for each of the ordered failure times for the entire data set. Going back to our maternal morbidity data example, we want to compare the Kaplan–Meier curves for the occurrence of hemorrhage for two women groups: women’s age at marriage  $\leq 14$  years versus  $\geq 15$  years. Twenty percent of women who married at  $\leq 14$  years experienced the event, compared to 13.9% whose age at marriage is  $\geq 15$  years. We used SPSS to perform the log-rank test. SAS codes for the log-rank test are given in Sect. 4. Figure 15.6 presents the survival and hazard function curves by age at marriage. By looking at the graphs (Fig. 15.6), you can see that the survival probability for women whose age at marriage is  $\leq 14$  years is lower at all time points than for those who got married at  $\geq 15$  years. The reverse is true for the cumulative hazard graph (Table 15.3).



**Fig. 15.6** The survival and hazard function curves by age at marriage

**Table 15.3** Distribution of hemorrhage by age at marriage

agem2 \* Hemorrhage cross-tabulation

			Hemorrhage		Total	
			No	Yes		
agem2	$\leq 14$ year	Count	64	16	80	
		% within agem2	80.0%	20.0%	100.0%	
	$\geq 15$ year	Count	93	15	108	
		% within agem2	86.1%	13.9%	100.0%	
Total		Count	157	31	188	
		% within agem2	83.5%	16.5%	100.0%	

**Table 15.4** Test of equality of survival distributions for the categories of age at marriage*Overall comparisons*

	Chi-square	df	Sig.
Log rank (Mantel-Cox)	1.323	1	0.250
Breslow (generalized Wilcoxon)	1.287	1	0.257
Tarone-Ware	1.304	1	0.254

However, if you look at the test result in Table 15.4, the log-rank test results were not statistically significant between the two categories of age at marriage. Also, conclusions drawn from two other tests, known as Breslow (Generalized Wilcoxon) and Tarone–Ware, are similar to that of the log-rank test. Notably, reliable results are provided by the above tests only when the sample sizes are reasonably large. Unfortunately, there are no universally acceptable guidelines regarding which test to use in a specific situation. The Breslow–Wilcoxon method tends to put greater weight on events occurring early in the study period. Hence, the results can be misleading if a high proportion of subjects are censored early. The log-rank test, in contrast, gives equal weight to all events. If the assumption of proportional hazards holds true, then the log-rank test is more powerful of the two tests. Proportional hazard model implies that the ratio of two hazard functions (events per time) remains constant over time. The generalized Wilcoxon tests are useful in detecting differences in the two survival curves during the early time periods. In contrast, the log-rank test is better for detecting differences between the curves at the latter periods. In the event that two Kaplan–Meier survival curves cross each other, it indicates a non-proportional hazard. In such cases, other tests, such as Peto–Peto Prentice, Tarone–Ware, or Fleming–Harrington tests, may be needed instead of a log-rank test.

## 6 Other Survival Analysis Methods

Survival data has two main characteristics: (1) time—The follow-up time of each individual being followed and (2) censoring variable—Whether or not each individual suffers from the variable of interest during the study period. Most of the survival data have additional characteristics, in addition to the above-mentioned two variables. These include demographic variables, such as age, gender, socioeconomic status, education, and behavioral variables, which may affect the event of interest. Such variables can be used as covariates (also called explanatory variables, confounders, risk factors, independent variables, etc.) in explaining the response (dependent) variable. If you adjust for these potential explanatory variables, the results of survival times between groups should be less biased and more precise. One such technique is the widely used multiplicative hazards model, often called the **proportional hazards model**.

### 6.1 The Cox Proportional Hazards Model

The proportional hazard model is a type of survival models to analyze time-to-event data. Cox (1972) proposed this model, which can incorporate censored data, hence the name Cox Proportional Hazards (PH) Model [10]. Survival models relate the time before some event occurs to one or more covariates that may be associated with time. For example, age at marriage may affect the diseases affecting women—thus, it may affect maternal morbidity hazard rate. Another type of survival models is known as accelerated failure time model. It does not exhibit proportional hazards. The accelerated failure time model involves a situation in which an event's biological or mechanical life history is accelerated (or decelerated). The Cox PH model is written in terms of the hazard model formula below:

$$h(t, \mathbf{X}) = h_0(t) e^{\sum_{i=1}^p \beta_i X_i},$$

where  $\mathbf{X} = (X_1, X_2, \dots, X_p)$  are explanatory or predictor variables. In the Cox model, the hazard at time  $t$  is the product of two quantities. The first one,  $h_0(t)$ , is the baseline hazard function. The second quantity  $\exp\left(\sum_{i=1}^p \beta_i X_i\right)$  is the exponential expression to the linear sum of  $\beta_i X_i$ , where the sum is over the  $p$  explanatory variables.

**The Proportional Hazards (PH) Assumption:** In the above Cox PH model, the baseline hazard is a function of  $t$  but does not involve the  $X$ 's.

Here, the exponential expression consists of the  $X$ 's but does not include  $t$ . The  $X$ 's, therefore, are called **time-independent covariates**. However, if the covariates ( $X$ 's) are time-dependent, meaning that if the covariate values change over time, then the proportional hazard (PH) assumption is no longer satisfied. In that case, you should use an extended Cox PH model.

**Example** Let us take the example of maternal morbidity data to illustrate the Cox PH model. The covariates used include: age at marriage, antenatal care visits to a clinic, and economic status. Figure 15.7 presents the portion of outputs from the Cox PH model using SAS PROC PHREG. Let us analyze the data output of Fig. 15.7.

**Model Fit Statistics** They are typically used for comparison and selection among the multiple models. In this example, we have only one model, so there is no other model to compare. However, SAS will display the fit statistics of a model with no predictors by default. Adding three covariates as predictors (age at marriage, ante-

Model Fit Statistics			
Criterion	Without Covariates	With Covariates	
-2 LOG L	319.051	295.290	
AIC	319.051	303.290	
SBC	319.051	309.026	

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	23.7615	4	<.0001
Score	22.0023	4	0.0002
Wald	15.9781	4	0.0030

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
agem2	1	0.3405	0.5595
visit	1	13.0334	0.0003
eco	2	1.0991	0.5772

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
agem2	15+	1	-0.22196	0.38036	0.3405	0.5595	0.801	agem2 15+
visit	No	1	-1.98374	0.54949	13.0334	0.0003	0.138	visit No
eco	Hlgh	1	0.44864	0.42816	1.0980	0.2947	1.566	eco Hlgh
eco	Low	1	0.11584	0.48863	0.0562	0.8126	1.123	eco Low

Fig. 15.7 Outputs from the Cox PH model using SAS PROC PHREG

natal care, and economic status) improves the model's fit, as all three statistics decreased.

**Testing Global Null Hypothesis**  $\beta$  (beta) = 0; it evaluates the null hypothesis that all coefficients in the model are 0. Alternative hypothesis: At least one coefficient is not zero. These tests are asymptotically equivalent, but they may differ in smaller samples, in which case **the likelihood ratio test** is preferable. Here, at least one of the regression coefficients is significantly different from 0 because the  $p$ -value (see the Pr > ChiSq column) for all three test statistics such as Likelihood ratio, Score, and Wald decreased to <0.0001, 0.0002, and 0.003, respectively.

**Analysis of Maximum Likelihood Estimates** You may find the following data outputs (Fig. 15.7): model coefficients in the parameter estimates column; tests of significance in the Chi-square column and the Pr > ChiSq column, and exponentiated coefficient as the hazard ratio under the hazard ratio column. It appears that women with a marriage age of 15+ years have a hazard ratio of 0.801 decrease in the hazard rate compared to those whose age at marriage is 14 years or below. However, this decrease is not statistically significant because  $p > 0.05$  ( $p$ -value = 0.5595). Economic status had three categories, with "Medium" being the reference. Coefficients of high economic status ( $p = 0.2947$ ) and low economic status ( $p = 0.8126$ ) are also not statistically significant. However, the coefficient for the visit (antenatal care visits) is statistically significant ( $p = 0.003$ ). The hazard ratio for pregnant women who did not visit a clinic for antenatal care is 0.138, which is reduced significantly compared to those who visited a clinic for antenatal care. However, this finding is surprising, as one would expect the opposite. We can only speculate the explanations as follows: in this sample population, probably the women who did not visit are from the low socioeconomic group, and they need to manage all household works, which make them more fit than women from higher socioeconomic groups. There could be other factors that may interact with the visit covariate. As our objective is the illustration of the Cox PH model, we did not proceed any further. Interested readers may read other papers to see a comprehensive analysis of this data set using different survival analysis models [11, 12].

## 7 Survival Analysis Using SAS and SPSS

For this chapter, all the illustrations are based on maternal morbidity data collected longitudinally. We used both SAS and SPSS for illustration purposes only. Of course, you can use other statistical software such as R and Stata.

In this section, the Kaplan–Meier analysis is performed using SAS. Interested readers may go through an excellent book for SAS written by Cody [13]. The following is the SAS code to open the SPSS data file in SAS and perform KM analysis, and we will briefly illustrate the code. First, we used the PROC IMPORT procedure to read the SPSS data into SAS. The following code chunk will read the data file and

store it in WORK, which is SAS default content. The second line specifies the location and the file name, and the code in the third line defines that it is an SPSS data file.

```
PROC IMPORT OUT= WORK.Amal20
  DATAFILE=  'C:\Users\16472\Desktop\AmalMitra\ThesisDataUWO_
GROUP20.sav'
  DBMS=SAV REPLACE;
RUN;
```

The following SAS codes from the PROC lifetest procedure are used for the KM method to produce Figs. 15.4 and 15.5. The code in the first line defines the data file also to show the number at risk in Fig. 15.4 and stores all the outputs in outKM object. In the second line, time is the default and then Hemodur is the time-to-event variable in our data set; after the asterick, f1v6a\_1 is the censoring indicator variable in the data, 0 within the bracket is to denote censored time.

```
proc lifetest data=Amal20 atrisk outs=outKM;
  time Hemodur*f1v6a_1(0);
run;
```

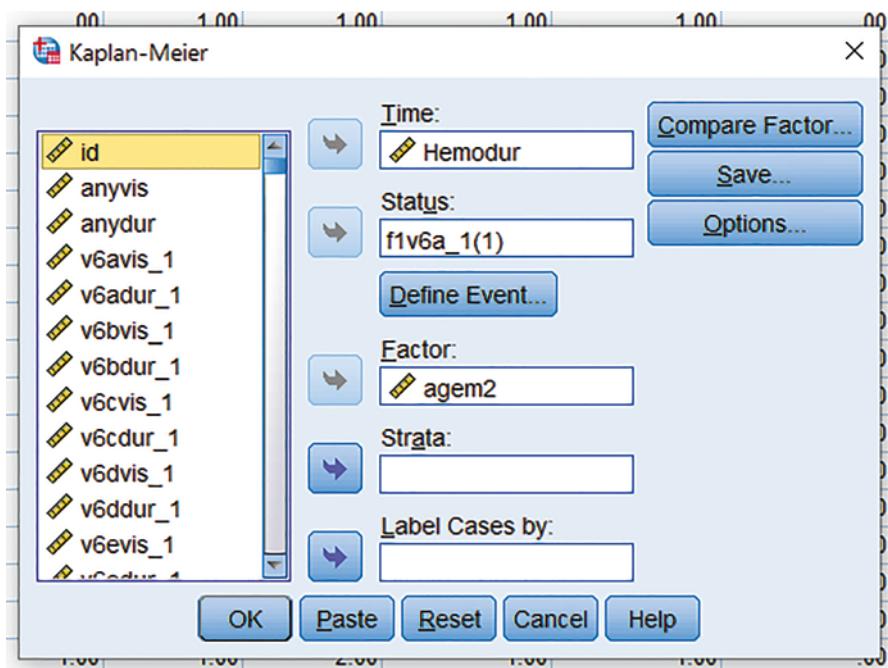
One needs to use the following SAS code to perform a log-rank test. This code chunk differs from the previous one with one extra row **strata agem2** to perform a log-rank test between the two age categories at marriage.

```
proc lifetest data=Amal20 atrisk outs=outLRT;
  strata agem2;
  time Hemodur*f1v6a_1(0);
run;
```

However, the log-rank test results in Table 15.4 and Fig. 15.6 are performed by using SPSS menu as follows: **Analyse → Survival → Kaplan–Meier** (Fig. 15.8). Here, this data set is presented in SPSS file format (.sav extension). In SPSS, first, you need to open the data file from the **File → Open → Data** menu, as shown in Fig. 15.8. Then, from the popup window, select the data file and click on Open; it will open, and you are ready to work.

## 8 Further Practice

1. Mention four examples of “lifetimes (survival times)” in epidemiology practice.
2. State the characteristics of a lifetime distribution.
3. State the properties of the survival function of a lifetime distribution.



**Fig. 15.8** Log-rank test using SPSS

4. Explain why the hazard function may be preferred to the survival function in exploring the behavior of a lifetime distribution.
5. Express the survival function  $S(t)$  and the probability density function  $f(t)$  in terms of the hazard function  $h(t)$ .
6. Consider the following right-censored data. No left truncation exists, and a plus sign (+) indicates right-censored observations.
  - (a) 2, 10+, 10+, 9, 10+, 20+, 7, 12, 9, 17, 21, 23, 27+.
  - (b) Use SAS or SPSS to obtain Kaplan–Meier estimates, as shown in Fig. 15.4, and plot the survival function, as in Fig. 15.5.
7. Also, prepare a plot of the Nelson–Aalen cumulative hazard function estimates using the data in Question 6.
8. Further, calculate  $\hat{S}^{KM}$ , the Kaplan–Meier estimate of  $S(9)$ , using the data from Question 6.
9. The following data set includes four variables for 30 pregnant women imitating this chapter's maternal morbidity data example. Suppose the variable **Event** represents the occurrence of hemorrhage for the first time during the antenatal period. The variable **Time** is the time to hemorrhage, the variable **Agem** is the age at marriage (0 = less than or equal to 15 years and 1 = 16 years or more), and **Eco** represents the economic status of the women (0 = low and 1 = high).

- (a) Time: 2.39 2.61 2.94 2.98 3.07 3.2 3.24 3.53 3.66 3.8 10.35 2.39 2.75 4.02  
4.19 4.22 4.29 4.61 5.07 5.47 5.76 5.83 5.96 6.02 6.35 6.58 6.61 6.98  
7.04 7.04
- (b) Event: 0 0 0 0 0 0 0 0 0 0 0 1
- (c) Agem: 0 0 0 1 0 1 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 1 0 0 0
- (d) Eco: 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 1 0 0 1 1 1 0 1 1 0

10. Perform the following tasks:

- (a) Read the data either in SAS or SPSS.
- (b) Use SAS or SPSS to obtain Kaplan–Meier estimates, as shown in Fig. 15.4, and plot the survival function, as in Fig. 15.5.
- (c) Evaluate separately whether or not Kaplan–Meier curves for two categories of age at marriage and categories of economic status are statistically equivalent.
- (d) Fit a cox proportional hazards model using age at marriage and economic status as the covariates. Interpret your findings.
- (e) Prepare a plot of the Nelson–Aalen cumulative hazard function estimates using the data in Question 6.

### Answer Keys

1. (a) Time to the event (death) after detection of ovarian epithelial carcinoma.  
(b) Time to the event (both relapse and death) after the bone marrow transplantation.  
(c) Time to complications (e.g., hemorrhage, edema, excessive vomiting, fits/convulsion, and delivery complications) during the antenatal period.  
(d) Time to the development of diabetic retinopathy from the time of diagnosis of diabetes.
2. A typical feature of lifetime distribution (time to event) is that we may not be able to observe the occurrence of the event (e.g., death) for all the patients by the end of the observation period. Therefore, some patients' actual event occurrence times (or survival times) are unknown. In other words, some of the data is censored. Most survival analysis methods must consider the problem of censoring in the analysis to allow for valid inferences. Furthermore, the distribution of survival times is usually skewed, limiting analysis methods that require the assumption of normally distributed data. For the analysis of survival (or time-to-event) data, it is necessary to consider the occurrence of the event (outcome) of interest, as well as when (the time) that event occurred.
3. (a) They are nonincreasing; that is, they head downward as  $t$  increases;  
(b) At time  $t = 0$ ,  $S(t) = S(0) = 1$ ; that is, at the start of the study, since no one experienced the event, the probability of surviving past time 0 is one;  
(c) At time  $t = \infty$ ,  $S(t) = S(\infty) = 0$ ; theoretically, if the study period increased to infinity, nobody would eventually survive, so the survival curve must fall to zero.

4. (a) It is a measure of instantaneous potential, whereas a survival curve is a cumulative measure over time.  
 (b) It can be used to identify a specific model form (e.g., an exponential, a Weibull, or a lognormal curve that fits the Data).  
 (c) The survival models are commonly written in terms of the hazard function.
  5. These functions are mathematically related as below:
- $S(t) = \exp\{-H(t)\}$ , where  $H(t)$  is the integral of  $h(t)$  from 0 to time  $t$   
 $F(t) = 1 - \exp\{-H(t)\}$   
 $f(t) = h(t) \exp\{-H(t)\} = h(t) S(t)$
6. Use SAS or SPSS to obtain Kaplan–Meier estimates, as shown in Fig. 15.4, and plot the survival function, as in Fig. 15.5.
  7. Use SAS for this question. First, prepare a SAS data file using the data from Question 6 and use the following SAS codes.

```
ods output ProductLimitEstimates = ple;
proc lifetest data=Q6.Q6a nelson outs=outNel;
time Time*Censor(0);
run;

proc sgplot data = ple;
series x = Time y = CumHaz;
run;
```

8. From the survival table of SPSS output of Question 6, the estimate is 0.692.
9. and 10. Follow the chapter—Figs. 15.4 and 15.5

**Acknowledgments** Thanks to Dr. Halida Hanum Akhter, Director, BIRPERHT, for providing the data.

## References

1. Putter H, Fiocco M, Geskus RB. Tutorial in biostatistics: competing risks and multi-state models. Stat Med. 2007;26:2389–430. <https://doi.org/10.1002/sim.2712>.
2. Akhter HA, Chowdhury MEEK, Sen A. A cross-sectional study on maternal morbidity in Bangladesh. Bangladesh Institute of Research for Health and Technologies (BIRPERHT); 1996.
3. Kalbfleisch JD, Prentice RL. The statistical analysis of failure time data. New York: Wiley; 2002.
4. Klein JP, Moeschberger ML. Survival analysis: techniques for censored and truncated data. 2nd ed. New York: Springer-Verlag; 2003.
5. Kleinbaum DG, Klein M. Survival analysis: a self-learning text. 3rd ed. New York: Springer; 2012.
6. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53:457–81.

7. Nelson W. Theory and applications of hazard plotting for censored failure data. *Technometrics*. 1972;14:945–65.
8. Aalen O. Nonparametric inference for a family of counting processes. *Ann Stat*. 1978;6:701–26.
9. Borgan Ø. Nelson-Aalen estimator. Available at: <https://www.medicine.mcgill.ca/epidemiology/hanley/c609/material/NelsonAalenEstimator.pdf>. Accessed 20 May 2023.
10. LaMorte WW. Cox proportional hazards regression analysis. 2016. Available at: [https://sphweb.bumc.bu.edu/otl/mpb-modules/bs/bs704\\_survival/BS704\\_Survival6.html](https://sphweb.bumc.bu.edu/otl/mpb-modules/bs/bs704_survival/BS704_Survival6.html). Accessed 20 May 2023.
11. Chowdhury RI, Islam MA. Prediction of risks of sequence of events using multistage proportional hazards model: a marginal-conditional modelling approach. *Stat Methods Appl*. 2020;29:141–71.
12. Islam MA, Chowdhury RI, Chakraborty N, Bari W. A multistage model for maternal morbidity during antenatal, delivery and postpartum periods. *Stat Med*. 2004;23:137–58.
13. Cody R. Learning SAS by example--a programmer's guide. Cary: SAS Publishing; 2018.

# Chapter 16

## Systematic Review and Meta-Analysis: Evidence-Based Decision-Making in Public Health



Aliyar Cyrus Fouladkhah and Minoo Bagheri

### Learning Objectives

- Assimilating the importance of evidence-based decision-making in public health
- Learning various biases, complexities, and limitations for the conduct of systematic reviews and meta-analyses studies
- Gaining knowledge about analytical approaches for the conduct of meta-analysis

## 1 Introduction: Evidence-Based Decision-Making

There are major differences between advocacy and analysis for drawing a conclusion from the literature. If one is an advocate, there is a particular cause or policy in mind. Using that predetermined notion, an advocate collects information from the literature to support the particular policy or cause [1]. For example, for someone representing an organization that manufactures conjugated linolenic acid supplements, an advocate could identify studies that highlight the positive impact of conjugated linolenic acid supplements on health. Of course, this approach may not provide the complete picture of the health effects of the supplement as studies that show no positive outcome or those that may show negative health consequences.

---

A. C. Fouladkhah (✉)

Public Health Microbiology Laboratory, Tennessee State University, Nashville, TN, USA

Public Health Microbiology Foundation™, Nashville, TN, USA

e-mail: [aliyar.fouladkhah@aya.yale.edu](mailto:aliyar.fouladkhah@aya.yale.edu)

M. Bagheri

Department of Cardiovascular Medicine, Vanderbilt University Medical Center,  
Nashville, TN, USA

e-mail: [minoo.bagheri@vumc.org](mailto:minoo.bagheri@vumc.org)

associated with the intake of conjugated linolenic acid might not be collected by an advocate. In contrast to this approach, evidence-based decision-making is derived from a systematic literature analysis [2]. An analyst considers negative and positive results and those that do not illustrate significant differences. An analyst could make an evidence-based decision by considering important topics such as publication bias and *winner's curse*, using inclusion and exclusion criteria for evaluating existing studies, and using analytical tools.

Evidence-based decision-making starts with a review of the literature. For identifying the existing studies, various search engines could be utilized. Although the use of common search engines such as Google Scholar could lead to the identification of many existing studies, the use of specialized search engines could broaden the impact of the literature review efforts. PubMed is a free search engine that is maintained by the United States National Library of Medicine and the National Institute of Health that could be considered the best source for the identification of medical and public health peer-reviewed research articles. Many other databases could be of great help and importance as well. For example, AGRICOLA is a search engine maintained by the National Agriculture Library and the United States Department of Agriculture and could be a great resource for identifying nutrition-and food system-related articles. AGRIS is another helpful search engine maintained by the United Nations Food and Agricultural Organization that could be used to identify literature associated with nutrition, public health, and global food security. AGRIS is also a great resource for searching research articles that are published in a language other than English. Ovid, Web of Science, Cochrane library, and Scopus databases are also very common resources for obtaining information for systematic review and meta-analysis. A successful literature review requires the identification of appropriate keywords and maintaining a search record. More complex search efforts could be completed using Boolean operators (such as and, or, not, or and not) as conjugations to combine and exclude keywords [3]. Using these Boolean operators could make the review and evaluation of search records more manageable for filtering studies that are irrelevant to the search topic.

## 2 Limitations and Important Considerations for Systematic Review and Meta-Analysis

A successful and comprehensive literature review requires identifying all types of studies associated with a search topic, including the material often called "grey literature." In contrast to widely disseminated published literature, grey publications are typically not peer-reviewed and may contain results of studies that are not statistically significant. These negative-outcome studies are very important in the control of publication bias, as further discussed below. The unpublished grey literature could also be a good source of pilot projects that have very small sample size due to high cost or difficulty in recruiting subjects and studies that are highly innovative in nature [4–7].

### **Box 16.1 Evidence-Based Decision Making in Public Health**

An advocate has a particular cause or policy in mind and based on that pre-determined notion, collects information from the literature. This could lead to major biases in decision making based on information in the literature. In contrast, an analyst uses a systematic procedure to collect information from the literature, conducts quality assessment, considers both positive and negative results, and after controlling biases could summarize and interpret the results in a systematic review and/or meta-analysis. This approach leads to evidence-based decisions and could be completed by individuals with sufficient training who are intelligent consumer of the literature.

Notably, grey publications such as annual reports, project reports, working papers, government publications, and white papers might not be peer-reviewed, and including them in a systematic review should be done cautiously after carefully considering the scientific merit of the studies. Thus, the inclusion and exclusion of grey publications should be carefully considered. Some researchers might place a very heavy emphasis on the grey literature, as an example, Piggott-McKellar et al. completed a systematic review of only the grey literature associated with climate change [8]. In certain fields of study, information discussed in patents could also be considered a grey publication and be used in a systematic review [9]. In those cases, it is important to ensure that information derived from patents is discussed so that it does not violate the patent and copyright laws.

A very important topic for consideration during systematic review is publication bias which should not be confused with biases for selecting samples and participants in epidemiological studies. According to “the dictionary of Epidemiology,” [10] publication bias is defined as *“an editorial predilection for publishing particular findings, e.g., positive results, which leads to the failure of authors to submit negative findings for publication.”* This is particularly common for researchers who use analytical skills for drawing associations from large data sets and may not express interest in further pursuing a hypothesis that does not have a statistically significant difference. As apparent in the definition, publication bias could also occur systematically by journal editors and peer-reviewers by not showing interest in the publication of studies with negative results [11]. The lack of inclusion of negative results in a systematic review and meta-analysis could substantially affect the outcome of the analysis. Some studies indicate that as high as 45% of an observed association could be attributed to publication bias [11, 12]. As such, when a negative study is identified, it is very important to give appropriate weight. Discussion about publication bias could also be an important part of interpreting both systematic reviews and meta-analyses.

When conducting and interpreting systematic reviews and meta-analyses, it is also important to consider the *winner’s curse* effect in research publications. The verbiage *winner’s curse* is derived from competitive auctions when a “winner” of an item listed in an auction might have overpaid for an item. In scientific settings, a “winner,” i.e., the researcher who first identifies a novel association, might trumpet

important results that in further studies might illustrate less association or no significance. Follow-up and confirmatory studies to the original finding, since the nature of exposure and outcome is now determined based on the original work, might have a larger sample size and thus better estimate the impact of association on the outcome compared to the original study [11]. This could be of concern for genomic research, health policy studies, as well as public health randomized and observational investigations. As such, consideration of the *winner's curse* could be an important aspect of conducting and interpreting systematic review and meta-analysis. As discussed later in this chapter, conducting a sensitivity analysis for comparing the overall effect with and without the original study could be used as an analytical method for investigating the existence of the *winner's curse*. Careful examination of the normal quantile–quantile plot (QQ Plot) is also an important diagnostic tool to determine the potential effect of the *winner's curse* in a study.

For writing and summarizing material, it is important to mention that in recent years, there has been great progress in the use of artificial intelligence (AI) software such as ChatGPT. The use of these applications and ethical considerations of their utilization are currently in their infancy. Although these AI software applications could ultimately be of positive help for evidence-based decision-making, this technology should be used responsibly and ethically. When a researcher's work is utilized, proper credit should be given by properly citing the work. Similarly, material that is generated by AI technology or the work that has received assistance or derived from AI software applications requires proper citation of the work as well [13]. The use of existing software for determining similarity index reports, if used properly, could be of great help to examine if a written summary is properly giving credit to other researchers in the field, but at the current time, these software applications cannot detect the work generated by AI applications. However, emerging AI software is now becoming available to determine if an essay is generated by another AI [14, 15].

### **Box 16.2 Biases and Important Considerations for Systematic Review and Meta-Analysis**

For ensuring the success of an evidence-based systematic review and meta-analysis, it is important to ensure that in addition to peer-review studies, grey literature are considered as well. Although many forms of grey literature may not be peer-reviewed and their inclusion in the final systematic review and meta-analysis should be considered carefully after quality assessment, these sources of information could be of great importance to minimize the risk of publication bias. Publication bias is due to the fact that studies with negative results might not be selected by the editorial team of journals or by researchers' themselves for publication so published literature tend to be more about positive and significant results and lacks research with negative outcome. Consideration about *winner's curse* are additionally an important aspect of a successful systematic review and meta-analysis. This could be detected using diagnostic tools such as quantile-quantile plot (QQ Plot).

### 3 Overview of Step-by-Step Procedures

#### 3.1 Criteria for Inclusion and Exclusion of Studies

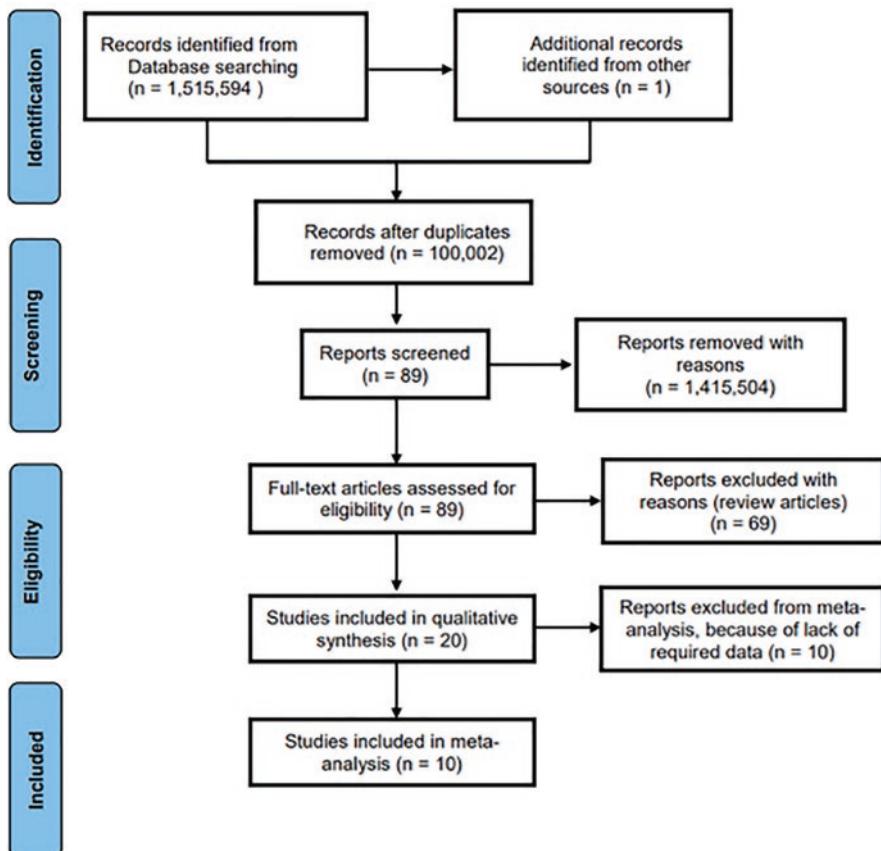
In addition to concerns associated with sample variability, a single study has limitations associated with the sample size and probability of false negatives in the outcome. Traditional narrative review articles also have major limitations since they are subjective in nature and prone to errors and biases [16]. In contrast, a successful systematic review and meta-analysis could lead to a precise estimation of treatment effects and a reduction of false-negative results probability.

However, the outcome of a systematic review and meta-analysis could be compromised if “raw materials” are derived from inadequate methodological quality.

A systematic review and meta-analysis should be very carefully planned. In addition to important information discussed earlier for selecting studies, it is important to have robust, transparent, and repeatable inclusion and exclusion criteria and quality assessment procedures [17]. The results of the quality assessment activities are usually summarized in the PRISMA flow chart. Figure 16.1 is an example of a PRISMA flow chart for studying racial disparities associated with respiratory infections in the US’ children [18]. The researcher could generate this flow chart or a template of a diagram that could be obtained from reputable agencies’ websites such as Cochrane Collaboration and the Cochrane Library [19].

Systematic review and meta-analysis could be completed for both randomized and observational studies. To ensure that biases are controlled in the final analysis, it is important to include robust inclusion and exclusion criteria for the quality assessment of the studies. The inclusion and exclusion criteria could vary depending on the type of studies and outcomes selected and the extent and availability of literature associated with the meta-analysis topic. These inclusion/exclusion criteria would need to be predefined. An example of a well-defined *a priori* inclusion/exclusion criteria could be accessed in the methods section of the study of Desai et al. [20]. In their study, the criteria were applied to 751 articles for the selection of 35 studies in their final systematic review.

Although the various methodologies could be applied for developing successful inclusion and exclusion criteria, Meline explains a six-step process for selecting studies for a systematic review [21] that includes the following: (i) applying the exclusion/inclusion criteria to abstracts and titles of the identified studies, (ii) elimination of studies that do not meet  $\geq 1$  of exclusion criteria, (iii) obtaining the full-text of remaining studies, (iv) further evaluation of studies’ full-text for inclusion and exclusion, (v) including studies that meet all inclusion criteria and do not meet any of the exclusion criteria, (vi) excluding studies during systematic review only with justified reasons (such as quality concerns), and (vii) accepting all remaining studies for a systematic review.



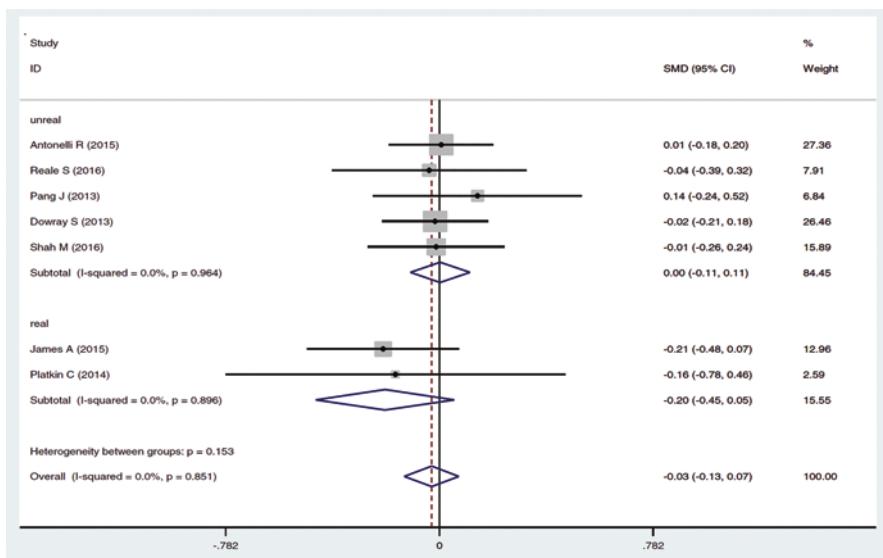
**Fig. 16.1** PRISMA flow chart for illustration of searched article and inclusion process [18] (Reprinted with permission from the authors)

### 3.2 Outcome Measures and Displaying Results

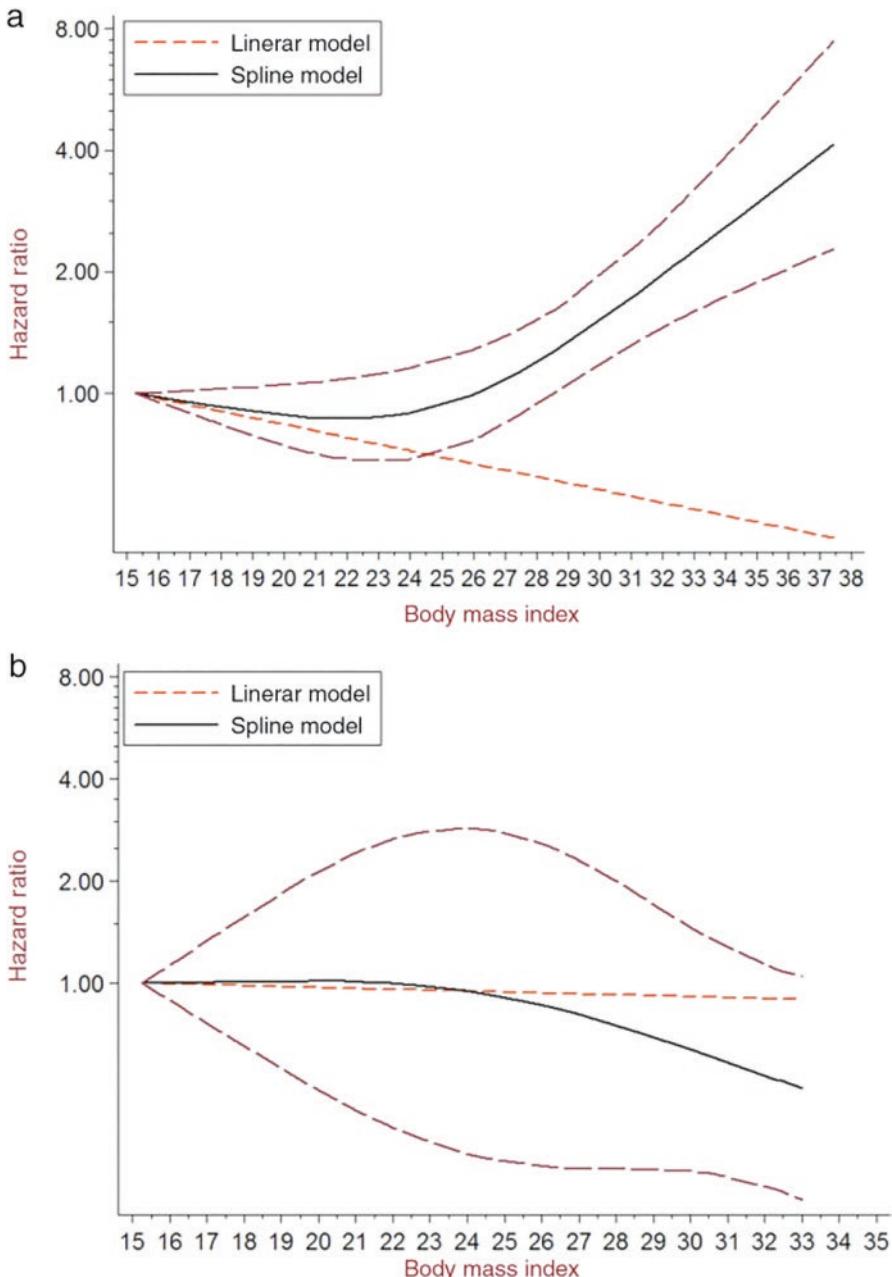
After a successful systematic review, choosing the outcome measure(s) that are intended to be studied is important. The type of outcome measures could be binary data, measures of associations, continuous variables, or time-to-event data. For choosing these outcome measures, it is important to ensure that these estimates are easy to interpret and are consistent across all selected studies. Additionally, it is important to consider the degree of heterogeneity of these outcome measures between selected studies. It is critical to consider the types of participants and interventions in each study to ensure that a formal meta-analysis can be considered for the data. A meta-analysis may have primary and secondary outcomes, and the analysis could be done for the main data and for the sub-groups [22]. Once these outcome measures are identified, they would need to be extracted from each study with

the sample size associated with each outcome measurement [23]. Since a meta-analysis aims to calculate the overall or combined effect of several studies, instead of relying on simple mean, it utilizes weighted mean, with more weight given to some studies and less weight given to others based on the sample size of each experiment [24, 25]. Although these could be calculated manually or in an Excel spreadsheet, many analytical software applications are available for data management and meta-analysis. As mentioned earlier, Cochrane reviews are typically considered as the gold standard of quality for systematic review and meta-analysis and the use of software from the Cochrane Library is recommended [26]. Some software, including the above-mentioned Cochrane software, allows the users to prepare a protocol that assists in searching and selecting studies and data extraction [26, 27]. Software applications such as SPSS® or programming software applications such as SAS®, Stata®, R®, and Python®, just to name a few, are all capable of conducting a meta-analysis as well. Once the weighted means are calculated, results are typically illustrated in a Forest plot. Figure 16.2 is an example of a Forest plot completed for continuous variables.

However, the meta-analysis results do not necessarily need to be presented in a forest plot. As an example, Fig. 16.3 shows the summary of a dose–response meta-analysis. A very common and important decision for conducting a meta-analysis is choosing fixed or random effects models. The general assumption for a fixed effects model is that there is one true effect size across all included studies. In contrast, a random effects model allows the flexibility that the true effect could be variable between the studies [24, 25, 30, 31].



**Fig. 16.2** Forest plot for the effect of physical activity label on calorie reduction by study setting [28] (Reprinted with permission from the authors)



**Fig. 16.3** Non-linear dose-response plots on the relationship between body mass index (BMI) and mortality in stroke patients. The association of BMI with all-cause mortality (a). The association of BMI with stroke specific mortality (b). Continuous black and medium-dashed orange-red lines represent non-linear and linear plots. 95% confidence intervals are shown by long-dashed maroon lines. Vertical axes are based on the log-scale of the hazard ratios [29]. (Used with permission from the publisher; permission conveyed through Copyright Clearance Center, Inc.)

### 3.3 Validity of Meta-Analysis and Conduct of Sensitivity Analysis

Sensitivity analysis could be defined as “*a method to determine the robustness of an assessment by examining the extent to which results are affected by changes in methods, models, values of unmeasured variables, or assumptions [32, 33]*.” The sensitivity analysis could be either quantitative or qualitative. For example, for formulating a research question before initiation of a literature review, one could ask “what if” questions (qualitative sensitivity analysis) to entertain various variations of research questions since this step could determine the outcome and analytics needed for a successful systematic review and meta-analysis [34]. However, sensitivity analysis typically refers to quantitative and analytical approaches to examine the robustness of the study. Sensitivity analysis is an excellent option for the detection of publication bias, as discussed earlier. Some studies suggest conducting worse-case meta-analytic point estimates by conducting a standard meta-analysis only for nonsignificant and negative studies to explore the impact of these potentially underrepresented studies on the overall meta-analysis [35]. Funnel and QQ plots are also very common and impactful diagnostic tools to examine the validity of meta-analyses [32]. Conducting these diagnostics tests and sensitivity analysis could be of great importance to determine the effects of outliers, missing data, publication bias, *winner’s curse*, or the existence of baseline imbalances on a meta-analysis’s internal and external validity [36]. These are also very important to determine the unexplained variations among various studies (heterogeneity) that could significantly threaten a meta-analysis’s internal and external validity [37].

Overall, a meta-analysis could be evaluated and reviewed for merit and significance by reviewing study questions (clear articulation of objectives, relevance to public health), quality of literature search, data abstraction procedure, evaluation of results and graphical displays, control measures for preventing or minimizing the publication bias, applicability of final results, and the funding source [38, 39]. These sections should be carefully reviewed for choice of analytical methods, risk of biases, inconsistencies, imprecision, indirectness, reporting biases, and confidence in the estimates [40].

## 4 Practical Examples and Interpretation of Relevant Studies

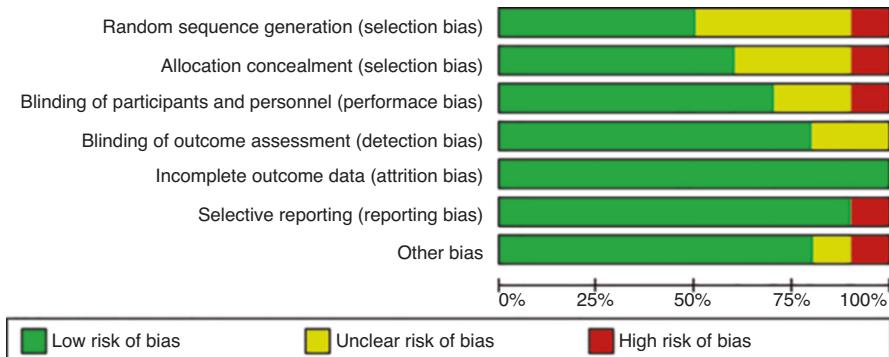
### 4.1 Example 1—Physical Activity Equivalent Labeling Versus Calorie Labeling

Several strategies are used in different countries to reduce the rates of overweight and obesity. One of these approaches includes nutrition labeling, a policy in which information like food calories on menus (calorie labeling) is provided to consumers to help them choose healthy food choices. Although this policy contributes to

changes in consumers' food selection and eating behaviors, there are controversies revolving around the effectiveness of the approach as it seems that it inversely affects eating behaviors. Alternatively, a more effective type of nutrition labeling that has been recently proposed is called physical activity equivalent labeling which is the amount of physical activity needed to burn calories taken by consuming each food item. Seyedhamzeh et al. [28] designed a meta-analysis to determine the effectiveness of equivalent physical activity labeling to calorie-only labeling on any alteration in consumers' eating behavior. This study searched several databases including PubMed, Scopus, Web of Science, AGRIS, Cochrane library, and Google Scholar to identify any scientific paper published between January 2000 and October 2016. They reported the findings using either or both labeling methods. The reference lists of these articles were also examined to prevent missing information published in this area. No restriction was performed with respect to the language of the relevant articles. If any discrepancies occurred during data extraction, a consensus was reached to make a decision. The Cochrane assessment tool could not be used for quality assessment as most of the included studies were performed in non-real-world settings. Therefore, the authors designed their own tool and categorized the studies into three groups: high risk of bias, low risk of bias, and unclear. The degree of heterogeneity between studies was evaluated using the Q Cochrane test and  $I^2$  statistics. Heterogeneity was classified as low, moderate, or severe based on  $I^2$  values of less than 25%, between 25% and 50%, and over 50%, respectively. A fixed effects model was utilized to merge the mean and standard deviations of the included studies. The authors also used a fixed effects model to calculate a weighted mean difference accounting for any differences in calorie purchases between those exposed to physical activity and calorie-only labeling. The findings of this study showed that calorie ordering during food selection for consumers exposed to physical activity labeling rather than calories only slightly decreased, but this reduction was not statistically significant when physical activity was shown either in minutes [standardized mean difference (SMD): -0.03; 95% confidence interval (CI): -0.13 to 0.07] or in miles [SMD-0.02; 95%CI: -0.13 to 0.09]. The findings on sub-group analysis based on quality assessment revealed that in low risk of bias studies, calories ordered in physical activity label compared to calorie label were nonsignificantly decreased [SMD: -0.04; 95%CI: -0.15 to 0.07]. They also found that the impact of those exposed to physical activity labeling on calorie reduction compared to individuals exposed to calorie labeling demonstrated an average decrease of 65 calories in their food choices (Fig. 16.2). This study concluded that there was no significant difference regarding calories ordered by consumers exposed to physical activity labeling and calorie labeling.

## 4.2 Example 2—*BMI Impact on Stroke and All-Cause Mortality*

The “obesity paradox” is centered upon the protective role of fat in individuals with metabolic conditions. However, it was unclear if this effect, also known as reverse epidemiology, exists in all-cause mortality and mortality from a specific disease in patients suffering from that particular disease. Therefore, Bagheri et al. [29] designed a dose-response meta-analysis of the studies examining the relationship between body mass index (BMI) and mortality due to all causes and stroke-specific causes in individuals suffering from a stroke. PubMed, Ovid, and Scopus databases were systematically searched to identify these studies published in English before 7 July 2014, and a predefined protocol was followed to extract the studies. This protocol includes studies with (1) the adult population (aged >18 years), (2) the measurement of BMI as the exposure and mortality as the outcome in the patients suffering from stroke, (3) findings reported as the relative risk (RR) or hazard ratio (HR), and (4) reported number of cases and controls in each BMI classes. To include all possible data, some studies’ authors were contacted and asked for additional information. Two authors independently extracted the data, and a third author resolved disagreements by consensus. The heterogeneity and the classification of the heterogeneity were assessed using the method explained earlier in Study 1. Based on the above protocol, the selected studies on (1) all-cause mortality were eight cohorts comprising 20,807 deaths of 95,651 stroke patients and (2) stroke-specific mortality were nine studies comprising 8087 deaths of 286,270 patients. The authors evaluated the quality of each study based on selection, comparability, and outcome using the Newcastle–Ottawa Scale. A two-stage hierarchical regression model performed the nonlinear dose–response association across all BMI categories. The spline transformations were used to estimate the dose–response relationship, while within- and between-study variances were taken into account. To assess this relationship, authors used random-effects dose–response models, which considered logarithms of HRs and CIs, as well as the number of deaths and participants across various BMI categories. Linearity in the potential relationship was assumed during the analysis. The findings of this study revealed a statistically significant nonlinear relationship ( $P < 0.0001$ ) between BMI and mortality due to all causes following stroke. For this analysis, the authors reported the correlation matrix estimate of 0.72 and the estimated between-study standard deviations (SDs) of 0.03 and 0.01. Results showed that BMI values below  $25 \text{ kg/m}^2$  had a protective effect on all-cause mortality rates, with the decreasing risk as BMI increased up to  $21 \text{ kg/m}^2$ , but the increasing risk with a steep slope at BMI levels above 23 (Fig. 16.3a). Also, the random-effects dose–response analysis showed a nonlinear relationship ( $P = 0.05$ ) between BMI and mortality specific to stroke (the estimate of the correlation matrix was reported to be  $-1$  and the estimated between studies SDs were 0.05 and 0.02) (Fig. 16.3b). Potential sources of heterogeneity were observed using the Q Cochrane test and  $I^2$  statistics. In addition, based on Egger’s regression test, there was not any publication bias among the included cohorts in



**Fig. 16.4** Risk of bias graph: review authors' judgements about each risk of bias item presented as percentages across all included studies. (Reprinted with permission from the author)

this meta-analysis. Based on the findings of this meta-analysis, an increase in BMI, an increased rate of all-cause mortality, and a decreased rate of mortality specific to stroke were observed. This showed that in patients suffering from stroke, different patterns were found in all-cause mortality and stroke-specific mortality with respect to BMI increase, indicating that there might be a paradox within the obesity paradox in this relationship.

### 4.3 Risk of Bias Assessment

In a meta-analysis, risk of bias assessment is an important step for quality control of the studies used for the analysis. In a recently published systematic review and meta-analysis of the safety and efficacy of post-eradication smallpox vaccine for the prevention of mpox (monkeypox), the investigators independently assessed the risk of bias in the included study using predetermined criteria, such as random sequence generation, allocation concealment; blinding of participants, personnel, and outcomes; incomplete outcome data; selective outcome reporting; and other sources of bias, in accordance with the methods recommended by The Cochrane Collaboration [41]. The results are presented in Fig. 16.4. The risk of bias was from low (green) to high (red) in the studies reviewed.

## 5 Further Practice

1. **What are the differences between advocacy and analysis in the context of evidence-based decision-making?** An advocate has a particular cause or policy in mind and based on that predetermined notion collects information from literature. This could lead to major biases in decision-making based on

information in the literature. In contrast, an analyst uses a systematic procedure to collect information from the literature, conducts quality assessment, considers both positive and negative results, and after controlling biases could summarize and interpret the results in a systematic review and/or meta-analysis. This approach leads to evidence-based decisions and could be completed by individuals with sufficient training who are intelligent consumers of the literature.

2. **What are PubMed, AGRICOLA, and AGRIS and what are the institutions that maintain these search engines?** PubMed is a free search engine that is maintained by the United States National Library of Medicine and the National Institute of Health that could be considered the best source for the identification of medical and public health peer-reviewed research articles. AGRICOLA is a search engine maintained by the National Agriculture Library and the United States Department of Agriculture and could be a great resource for identifying nutrition- and food system-related articles. AGRIS is another helpful search engine maintained by the United Nation's Food and Agricultural Organization that could be used to identify literature associated with nutrition, public health, and global food security. AGRIS is also a great resource for searching research articles that are published in a language other than English.
3. **What are Boolean operators and why are they important in a successful literature review?** Boolean operators (such as and, or, not, or and not) are conjunctions that could be used to combine or exclude keywords. The use of these Boolean operators could make the review and evaluation of search records more manageable for filtering studies that are not relevant to the search topic.
4. **What are grey publications and what is their importance in a successful systematic review?** A successful and comprehensive literature review requires the identification of all types of studies associated with a search topic including the material that is often called “grey literature.” In contrast to widely disseminated published literature, grey publications are typically not peer-reviewed and may contain results of studies that are not statistically significant. These negative-outcome studies are very important in the control of publication bias as further discussed below. The unpublished grey literature could also be a good source of pilot projects that have very small sample size due to high cost or difficulty in recruiting subjects and studies that are highly innovative in nature.
5. **What is publication bias in the context of systematic review and meta-analysis?** A very important topic for consideration during systematic review is publication bias which should not be confused with biases for the selection of samples and participants in epidemiological studies. According to “dictionary of epidemiology,” publication bias is defined as “*an editorial predilection for publishing particular findings, e.g., positive results, which leads to the failure of authors to submit negative findings for publication*” [10]. This is particularly common for researchers who use analytical skills for drawing associations from large data sets and may not express interest in further pursuing a hypothesis that does not have a statistically significant difference. As apparent in the definition, publication bias could also occur systematically by journal editors

and peer-reviewers by not showing interest in the publication of studies with negative results.

6. **What is the *winner's curse* effect in the context of systematic review and meta-analysis?** It is also important to carefully consider the *winner's curse* effect in research publications when conducting and interpreting systematic reviews and meta-analyses. The verbiage *winner's curse* is derived from competitive auctions when a “winner” of an item listed in an auction might have overpaid for an item. In scientific settings, a “winner,” i.e., the researcher that first identifies a novel association might trumpet important results that in further studies might illustrate less degree of association or no significance. Follow-up and confirmatory studies to the original finding, since the nature of exposure and outcome is now determined based on the original work, might have a larger sample size and thus better estimate impact of association on the outcome compared to the original study. This could be of concern for genomic research, health policy studies, as well as public health randomized and observational investigations. As such, consideration of the *winner's curse* could be an important aspect of the conduct and interpretation of systematic review and meta-analysis.
7. **What are the limitations of a single study and a traditional review article, and how these limitations are addressed in a systematic review and meta-analysis?** In addition to concerns associated with sample variability, a single study has limitations associated with the sample size and probability of false negatives in the outcome. Traditional narrative review articles also have major limitations since they are subjective in nature and prone to errors and biases. In contrast, a successful systematic review and meta-analysis could lead to a precise estimation of treatment effects, and a reduction of false-negative results probability. However, the outcome of a systematic review and meta-analysis could be compromised if “raw materials” are derived from inadequate methodological quality.
8. **What is a PRISMA flow chart?** The results of the quality assessment activities are usually summarized in the PRISMA flow chart. This flow chart that illustrates number of identified and number of selected studies could be generated by the researcher, or a template of a diagram could be obtained from reputable agencies’ websites such as Cochrane Collaboration and the Cochrane Library.
9. **What are typical outcome measures in meta-analyses and what are common software applications used for conducting meta-analysis?** Outcome measures could be binary data, measures of associations, continuous variables, or time-to-event data. For choosing these outcome measures, it is important to ensure that these estimates are easy to interpret and are consistent across all selected studies.

Cochrane reviews are typically considered as the gold standard of quality for systematic review and meta-analysis and the use of software from the Cochrane library is recommended. Some software including the Cochrane software allows the users to prepare a protocol that assists in searching and selecting

studies and data extraction. Software applications such as SPSS® or programming software applications such as SAS®, Stata®, R®, and Python®, just to name a few, are all capable of conducting a meta-analysis as well.

10. **What is sensitivity analysis and why is it important?** Sensitivity analysis could be defined as “*a method to determine the robustness of an assessment by examining the extent to which results are affected by changes in methods, models, values of unmeasured variables, or assumptions*” [32, 33]. The sensitivity analysis could be both quantitative and qualitative in nature. For example, for formulating a research question before initiation of a literature review, one could ask “what if” questions (qualitative sensitivity analysis) to entertain various variations of research questions since this step could determine the outcome and analytics needed for a successful systematic review and meta-analysis. However, sensitivity analysis typically refers to quantitative and analytical approaches to examine the robustness of the study. Sensitivity analysis is an excellent option for the detection of publication bias.

## References

1. Reid EJ. Understanding the word “advocacy”: context and use. In: Structuring the inquiry into advocacy, vol. 1. Urban Institute; 2000. p. 1–7.
2. Brownson RC, Gurney JG, Land GH. Evidence-based decision making in public health. *J Public Health Manag Pract.* 1999;5:86–97.
3. Jesson J, Matheson L, Lacey FM. Doing your literature review: traditional and systematic techniques. Los Angeles: SAGE; 2011.
4. Conn VS, Valentine JC, Cooper HM, Rantz MJ. Grey literature in meta-analyses. *Nurs Res.* 2003;52(4):256–61.
5. Adams J, Hillier-Brown FC, Moore HJ, Lake AA, Araujo-Soares V, White M, Summerbell C. Searching and synthesising “grey literature” and “grey information” in public health: critical reflections on three case studies. *Syst Rev.* 2016;5(1):1–11.
6. Hopewell S, McDonald S, Clarke MJ, Egger M. Grey literature in meta-analyses of randomized trials of health care interventions. *Cochrane Database Syst Rev.* 2007;2000(2):MR000010.
7. Saleh AA, Ratajeski MA, Bertolet M. Grey literature searching for health sciences systematic reviews: a prospective study of time spent and resources utilized. *Evid Based Libr Inf Pract.* 2014;9(3):28.
8. Piggott-McKellar AE, McNamara KE, Nunn PD, Watson JE. What are the barriers to successful community-based climate change adaptation? A review of grey literature. *Local Environ.* 2019;24(4):374–90.
9. Fouladkhah A, Berlin D, Bruntz D. High-sodium processed foods: public health burden and sodium reduction strategies for industry practitioners. *Food Rev Intl.* 2015;31(4):341–54.
10. Last JM, editor. A dictionary of epidemiology. Oxford: Oxford University Press; 1983. p. 12.
11. Thornton A, Lee P. Publication bias in meta-analysis: its causes and consequences. *J Clin Epidemiol.* 2000;53(2):207–16.
12. Tweedie RL, Scott DJ, Biggerstaff BJ, Mengersen KL. Bayesian meta-analysis, with application to studies of ETS and lung cancer. *Lung Cancer.* 1996;14:S171–94.
13. Lund BD, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? *Libr Hi Tech News.* 2023;40:26.
14. Meo SA, Talha M. Turnitin: is it a text matching or plagiarism detection tool? *Saudi J Anaesth.* 2019;13(Suppl 1):S48.

15. Fouladkhah A. Emerging open-access artificial intelligence applications: a peer-to-peer note to teaching colleagues in higher education. LinkedIn. 23 Feb 2023. [https://www.linkedin.com/posts/aliyarfouladkhah\\_artificialintelligence-highereducation-studentsevaluation-activity-7016170362358923264-X96h?utm\\_source=share&utm\\_medium=member\\_desktop](https://www.linkedin.com/posts/aliyarfouladkhah_artificialintelligence-highereducation-studentsevaluation-activity-7016170362358923264-X96h?utm_source=share&utm_medium=member_desktop).
16. Teagarden JR. Meta-analysis: whither narrative review? *Pharmacotherapy*. 1989;9(5):274–84.
17. Egger M, Smith GD, Phillips AN. Meta-analysis: principles and procedures. *BMJ*. 1997;315(7121):1533–7.
18. Jones EA, Mitra AK, Malone S. Racial disparities and common respiratory infectious diseases in children of the United States: a systematic review and meta-analysis. *Diseases*. 2023;11(1):23.
19. Clarke M. The Cochrane collaboration and the Cochrane library. *Otolaryngol Head Neck Surg*. 2007;137(4\_suppl):S52–4.
20. Desai MM, Stauffer BD, Feringa HH, Schreiner GC. Statistical models and patient predictors of readmission for acute myocardial infarction: a systematic review. *Circ Cardiovasc Qual Outcomes*. 2009;2(5):500–7.
21. Meline T. Selecting studies for systemic review: inclusion and exclusion criteria. *Contemp Issues Commun Sci Disord*. 2006;33(Spring):21–7.
22. Zwahlen M, Renehan A, Egger M. Meta-analysis in medical research: potentials and limitations. In: *Urologic oncology: seminars and original investigations*, vol. 26, No. 3. Elsevier; 2008. p. 320–9.
23. Sutton AJ, Abrams KR, Jones DR. An illustrated guide to the methods of meta-analysis. *J Eval Clin Pract*. 2001;7(2):135–48.
24. Borenstein M, Hedges LV, Higgins JP, Rothstein HR. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Res Synth Methods*. 2010;1(2):97–111.
25. Borenstein M, Hedges L, Rothstein H. Meta-analysis: fixed effect vs. random effects. *Meta-analysis.com*. 2007. p. 1–30.
26. Henderson LK, Craig JC, Willis NS, Tovey D, Webster AC. How to write a Cochrane systematic review. *Nephrology*. 2010;15(6):617–24.
27. Higgins JP, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. *Cochrane handbook for systematic reviews of interventions*. Wiley; 2019.
28. Seyedhamzeh S, Bagheri M, Keshtkar AA, Qorbani M, Viera AJ. Physical activity equivalent labeling vs. calorie labeling: a systematic review and meta-analysis. *Int J Behav Nutr Phys Act*. 2018;15:1–13.
29. Bagheri M, Speakman JR, Shabbidar S, Kazemi F, Djafarian K. A dose-response meta-analysis of the impact of body mass index on stroke and all-cause mortality in stroke patients: a paradox within a paradox. *Obes Rev*. 2015;16(5):416–23.
30. Nikolakopoulou A, Mavridis D, Salanti G. How to interpret meta-analysis models: fixed effect and random effects meta-analyses. *BMJ Ment Health*. 2014;17(2):64.
31. Hunter JE, Schmidt FL. Fixed effects vs. random effects meta-analysis models: implications for cumulative research knowledge. *Int J Sel Assess*. 2000;8(4):275–92.
32. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf*. 2006;15(5):291–303.
33. Thabane L, Mbuagbaw L, Zhang S, Samaan Z, Marcucci M, Ye C, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. *BMC Med Res Methodol*. 2013;13(1):1–12.
34. Cooper H, Hedges LV, Valentine JC, editors. *The handbook of research synthesis and meta-analysis*. New York: Russell Sage Foundation; 2019.
35. Mathur MB, VanderWeele TJ. Sensitivity analysis for publication bias in meta-analyses. *J R Stat Soc Ser C Appl Stat*. 2020;69(5):1091.
36. Liu Z, Yao Z, Li C, Liu X, Chen H, Gao C. A step-by-step guide to the systematic review and meta-analysis of diagnostic and prognostic test accuracy evaluations. *Br J Cancer*. 2013;108(11):2299–303.

37. Trikalinos TA, Balion CM, Coleman CI, Griffith L, Santaguida PL, Vandermeer B, Fu R. Chapter 8: Meta-analysis of test performance when there is a “gold standard”. *J Gen Intern Med.* 2012;27:56–66.
38. Russo MW. How to review a meta-analysis. *Gastroenterol Hepatol.* 2007;3(8):637.
39. Crowther M, Lim W, Crowther MA. Systematic review and meta-analysis methodology. *Blood.* 2010;116(17):3140–6.
40. Murad MH, Montori VM, Ioannidis JP, Jaeschke R, Devereaux PJ, Prasad K, et al. How to read a systematic review and meta-analysis and apply the results to patient care: users' guides to the medical literature. *JAMA.* 2014;312(2):171–9.
41. Malone SM, Mitra AK, Nwanne A, Onumah NA, et al. Safety and efficacy of post-eradication smallpox vaccine as an mpox vaccine: a systematic review with meta-analysis. *Int J Environ Res Public Health.* 2023;20:2963. <https://doi.org/10.3390/ijerph20042963>.

# Chapter 17

## Sample Size Estimation



Amal K. Mitra

### Learning Objectives

After completing this chapter, you will be able to:

- Understand the importance of sample size estimation
- Apply basic concepts in estimating sample size
- Use real-life examples
- Describe error rates, effect size, confidence, and power
- Calculate sample size for different epidemiologic studies
- Use G\*Power for sample size and power curve

## 1 Introduction

“One size doesn’t fit all” – the same is true for sample sizes. You shouldn’t use the same sample size that another researcher used for the research you propose. The two research on the same topic may differ in many ways: (1) the objectives of the two studies could be different; (2) the outcome measurements may or may not be the same; (3) the target population from which the sample is drawn could be entirely different; (4) the parameters that you are using could be different from the parameters used by the previous study, and so on. At some point in time, when I was reviewing a project proposal submitted to the institutional review board, the project proposal included 24 samples. When I questioned how they came up with such a sample size, the answer shocked me – “we followed most of the studies previously done on this topic; they all used sample sizes of a similar number.” My point is that

---

A. K. Mitra (✉)

Department of Epidemiology and Biostatistics, Jackson State University, Jackson, MS, USA  
e-mail: [amal.k.mitra@jsums.edu](mailto:amal.k.mitra@jsums.edu)

if someone does not follow certain rules in calculating sample size, and you follow him, you are committing the same mistake. Always calculate sample size based on certain rules that I will describe in this chapter unless you do your study with the entire population. Remember, even if two statisticians calculate sample size for you, the calculated sample sizes of the two people could differ because they probably did not use the same exact baseline data, same error rates, same effect size, and the same rate of dropout or attrition rate in their calculations.

This chapter will first describe these basic ingredients of calculating sample size. Then we will go through step-by-step sample size calculations for different study designs. Also, a free statistical software called G\*Power will be demonstrated for calculating a cross-sectional study's sample size and power curve.

## 2 Steps of Hypothesis Testing

In most studies, you have a target population from which you gather data using a sample. Then, based on the sample data, you want to make an inference about the population from which the sample has been drawn. In this section, we will find out how many people you would need for the sample to carry out a study to get enough power and adequate sample size in the data analysis. You definitely want to make sure that your analysis results are meaningful. In other words, you want to conduct required statistical analysis of certain outcome variables of the sample data to find out the true parameters in the population. Suppose you are testing the impact of an intervention in your study. In that case, you aim to identify if the variable of interest in the intervention group who had received the test drug (or an intervention) significantly differs from that in the control group who had not received the intervention.

### Real-Life Example

*Streptococcus pneumoniae* is the leading bacterial cause of community-acquired respiratory tract infections. Beginning in the 1990s, many *pneumococcal* isolates in the United States showed decreased susceptibility to penicillin and other commonly used antibiotics [1]. Fluoroquinolones (FQs) are a popular class of antibiotics used to treat various infections, such as respiratory tract infections, and include medications such as ciprofloxacin, levofloxacin, moxifloxacin, ofloxacin, and delafloxacin [2]. Suppose, you propose a double-blind, randomized, controlled, clinical trial with ciprofloxacin and levofloxacin for the treatment of laboratory-confirmed cases of pneumonia due to *pneumococcus* infection. In your study, you consider the patients who are getting levofloxacin as the treatment or test group (Group A) and those who are getting ciprofloxacin as the control group (Group B).

## 2.1 Hypothesis

**Null hypothesis ( $H_0$ )** Mean days of recovery from *pneumococcal pneumonia* patients treated with levofloxacin (Group A) and mean days of recovery from *pneumococcal pneumonia* patients treated with ciprofloxacin (Group B) are equal (or not different).  $\mu_1 = \mu_2$

A null hypothesis is also known as the hypothesis of no difference.

**Alternative hypothesis ( $H_A$ )** Mean days of recovery from *pneumococcal pneumonia* patients treated with levofloxacin (Group A) and mean days of recovery from *pneumococcal pneumonia* patients treated with ciprofloxacin (Group B) are not equal.  $\mu_1 \neq \mu_2$

An alternative hypothesis is a statement that holds true if the null hypothesis is rejected by data analysis. The alternative hypothesis is the same as your research hypothesis, which you wanted to prove in your research.

## 2.2 Steps of Hypothesis Testing

The following basic steps are key concepts in conducting a hypothesis test.

1. State the null ( $H_0$ ) and the alternative hypothesis ( $H_A$ )
2. Assumptions
3. Set error rates ( $\alpha$  and  $\beta$ ) and the effect size ( $d$ )
4. Estimate sample size ( $n$ )
5. Gather data
6. Analyze the data
7. Reach to a decision
8. Conclusion

### Assumptions

We make the following assumptions about the study.

- A simple random sample was drawn from the population
- Samples are independent
- Distribution of data either normal or nonnormal (based on the data) – appropriate statistical tests will be used based on the data distribution.

**Error Rates** Let us examine the  $2 \times 2$  contingency table (Table 17.1) to calculate error rates.

**Type I Error** Reject  $H_0$  when it is true in the population

Probability of Type I error =  $\alpha$

**Type II Error** Fail to reject  $H_0$  when it is false in the population

**Table 17.1** A  $2 \times 2$  contingency table

Decision	Null ( $H_0$ ) is true in the population	Null ( $H_0$ ) is false in the population
Reject null ( $H_0$ )	Type I error	Power
Fail to reject null ( $H_0$ )	Confidence	Type II error

Probability of Type II error =  $\beta$

In general,  $\alpha = 0.05$ . Why do we use  $\alpha = 0.05$ ? Because it is a tradition, a convention, and most journals (and researchers) accept this value.

### **What Does $p \leq 0.05$ Mean?**

Probability of committing an error by rejecting  $H_0$  when it is true. It is 5 in 100 or less.

In other words, we are 95% confident that  $H_0$  is not true. So, we reject  $H_0$  and accept  $H_A$ .

### **Statistical Power**

In general, we use  $\beta = 0.1$  or  $0.2$  (in most human studies)

If we use  $\beta = 0.1$ , Power =  $1 - \beta = 1 - 0.1 = 0.90$  (or 90%)

If we use  $\beta = 0.2$ , Power =  $1 - \beta = 1 - 0.2 = 0.80$  (or 80%)

So, if we increase  $\beta$ , power decreases.  $\beta$  and power are inversely proportional.

### **Effect size ( $d$ )**

It is an arbitrary number that we set before we conduct a study. How far from the null is worthwhile or acceptable? Effect size is the difference between the mean from the baseline and the mean that you will have after an intervention. In other words, it is the difference in mean (or other statistic) that you expect before and after the treatment or intervention. In this case, we are using an effect size for the mean difference; however, it can be applied to estimate a difference in other statistics, such as proportion.

The investigator decides the effect size based on the following:

1. The previous documents or doing a literature search.
2. If there is not much evidence in the literature, we use anecdotal data based on our past experience.
3. If no data are available, then we use our intelligent guess, maybe by consulting experts in the same field.
4. Sometimes, we need a pilot study to come up with some reasonable baseline data to compare.

### **Relation Between $\alpha$ , $\beta$ , Effect Size ( $d$ ), and Sample Size ( $n$ )**

Rule of thumb: If two of the four items ( $\alpha$ ,  $\beta$ ,  $d$ , and  $n$ ) are fixed, the other two vary inversely.

Please note the following:

- A general conception is that  $\alpha$  and  $\beta$  vary inversely. *It is true only when n and d are fixed.* In reality,  $n$  and  $d$  are not fixed for all the studies. We use the sample size based on several factors. We also use the effect size based on the type of the study. Therefore,  $\alpha$  and  $\beta$ , in reality, are not inversely proportional.
- If  $\alpha$  and  $d$  are fixed, increasing the value of  $\beta$  will decrease the sample size ( $n$ ). That tells us if we increase the error rate, the sample size will increase, and vice versa.
- If  $\beta$  increases, power  $(1 - \beta)$  decreases. Therefore, if  $\beta$  decreases, power increases. What happens to the sample size? If  $\beta$  decreases, the sample size increases. So, if you want to increase the power of the study, the sample size will be increased.
- If  $\alpha$  and  $\beta$  are fixed,  $n$  varies inversely with the effect size,  $d$ . So, if you increase the effect size (the worthwhile difference), the sample size is smaller. Please note that when you conduct research, you must find out what difference of the mean (or proportion or other estimates) is worthwhile for the population. Suppose you test a new drug for meningitis. How much difference in mean days of recovery or recovery rate (in proportion) will be good for the population if we consider the new drug effective? That difference in mean (or proportion) is the effect size. Do not increase the effect size inappropriately because you want to decrease the sample size. That is unethical research.

### ***Concept of p-Value***

In the literature, you may find data and tables corresponding to  $p$  or probability value. As stated earlier, scientists agree that if the probability of rejecting a null hypothesis when it is true in the population is 0.05 or less (5% or less), we can still reject the null and conclude that there is a significant difference between the observations. For the mean, we can say the two (or more) means in the population are significantly different. We also call it statistically significant.

$P$ -value represents the probability of rejecting the null when it is true. If the  $p$ -value is  $\leq 0.05$ , data are statistically significant. The  $p$ -value is a function of sample size. Larger the sample size, the smaller the  $p$ -value.

### ***Decision Rule Regarding p-Value***

- If  $p \leq 0.05$ , then reject  $H_0$
- If  $p > 0.05$ , then fail to reject  $H_0$

This is an either-or situation, reject or fail to reject, no matter how small (or large) the  $p$ -value is. Some literature suggests a sliding rule of  $p$ -value. Many statisticians do not use a sliding rule of  $p$ -value. However, you may find in the literature some statements as follows:

- “ $p = 0.065$ ; the results are almost significant”
- “ $p < 0.0001$ ; the results are highly significant”

Ideally, you should fix the value of “ $\alpha$ ” to calculate the sample size before conducting the research. After you have completed your research and analyzed the data, there is no good reason to say that “my data are almost significant or missed the significance level because of a small sample size.” These kinds of statements should be avoided because  $p$ -value is either significant or not significant.

### One-Tailed and Two-Tailed Test

- If the outcome goes in one direction only – you can do a one-tailed test.
- If the outcome is uncertain, i.e., it can go in either direction, you should do a two-tailed test.

### Attrition Rate

Attrition is the loss of the study samples. The attrition rate is not a dropout rate. Attrition may result from dropping out and other reasons, such as students moving out to other schools, graduating early, or being retained. A high attrition rate may lead to biased results.

In research, attrition rates may depend on:

- Type of research – e.g., in a cohort study, the attrition rate is generally high.
- Time period – more the time period of the study, more the attrition.
- Data source – attrition is generally higher in community-based studies compared to hospital-based studies.
- Length of the questionnaire – a questionnaire should not be too long. Time is important. A standard time to fill out a questionnaire is around 15 minutes. The respondent may not have enough time and patience to answer a lengthy questionnaire.
- Sensitive questions – one should not ask any sensitive questions in a questionnaire. However, because of the nature of the study, you may need to use private questions. It may need a reasonable explanation for using private information. People may not feel comfortable providing any private information.

It would help if you had an estimate of the dropout rate in research in advance. Suppose your calculated sample size is 200, and your estimated dropout rate is 20%. You must add 20% of the sample size to make the required sample size.

Total estimated sample size = calculated sample size + percent of dropout

Calculated sample size = 200

20% of 200 = 40

Total required sample size =  $200 + 40 = 240$ .

### How to Get Baseline Data

Getting reasonably sound baseline data is key to estimating an appropriate sample size. Generally, you should get baseline data by literature search. Now, many studies on a similar topic could produce published data. What data should you use? A few guidelines are provided below:

- Choosing published data:
  - Choose data that are relevant and most appropriate for your study population. Suppose you are studying “People’s knowledge and attitudes about organ donation in Kuwait.” You didn’t find any published reports on this topic from Kuwait. Your best option would be to use studies that were done in the neighboring countries.
  - If you get multiple articles, choose the most recent one.
  - Choose the results of a study design that is similar to yours.
- Get baseline data for all the major outcome variables that you are using in your proposed study.
- If there is no published literature, you can conduct a pilot study to generate data.
- In some cases, because of a lack of appropriate data, make a professional judgment regarding the baseline data to calculate the sample size.
- You can do an interim data analysis of your ongoing study and revisit the sample size.

## 3 Estimation of Sample Size

The sample size formula depends on the study design because the estimates are different in different studies. For a quick reference, you can use the estimates as shown in Table 17.2.

**Table 17.2** Common statistical parameters used for sample size estimation

Type of study	Parameters used
Cross-sectional study	Prevalence
Case-control study	Odds ratio
Cohort study	Relative risk
Randomized controlled trial (RCT)	
(a) Quantitative study	Mean and standard deviation
(b) Qualitative study	Proportions

### 3.1 Sample Size for a Cross-Sectional Study

A cross-sectional study is also called a prevalence study. Remember, in a cross-sectional study, you never get a true incidence because you gather information about existing cases, including ongoing and new cases. A cross-sectional study is descriptive; data are collected only once, and multiple outcomes can be studied.

#### **Sample Size Formula Used for a Cross-Sectional Study [3]**

$$n = \frac{Z^2 * P(1-P)}{d^2}$$

Here,  $n$  = sample size,  $P$  = prevalence of the variable of interest, and  $d$  = precision, meaning how close you are to the population's true value.

**Problem 1** Pregnant women's knowledge gaps about breastfeeding are a concern in developing and developed countries [4]. Suppose you wanted to conduct a cross-sectional study to understand correct knowledge about breastfeeding benefits among low-income first-time pregnant women in Mississippi. Upon literature search, suppose you found 25% of low-income women had correct breastfeeding knowledge. Calculate an appropriate sample size with 95% confidence and 4% precision to conduct a study. You also apprehend that approximately 20% of the population may be dropped out of the study.

#### **Data**

$$P = 25\% \text{ or } 0.25$$

$$1 - P = 1 - 0.25 = 0.75$$

$$\text{For 95\% confidence, } \alpha \text{ (alpha)} = 1 - 0.95 = 0.05$$

$Z$  value for  $\alpha = 0.05$ , and for a two-tailed test is 1.96 (a constant, which can be gathered from a  $Z$ -table of a statistics textbook).

$$d \text{ (precision)} = 4\% = 0.04$$

Dropout = 20% (which should be added after calculating the sample size).

#### **Calculations**

$$n = \frac{Z^2 * P(1-P)}{d^2}$$

$$n = \frac{(1.96)^2 * (0.25)(0.75)}{(0.04)^2}$$

$$n = 450$$

Add 20% of 450 = 90

The required sample size for the study,  $n = 450 + 90 = 540$

### 3.2 Sample Size for a Quantitative Study (Using Mean and Standard Deviation)

Use the following formula [5] if you are given mean and standard deviation (or standard error) to calculate the sample size:

$$n = \frac{2\sigma^2}{(\mu_1 - \mu_2)^2} \times f(\alpha, \beta)$$

Here,  $n$  = sample size,  $\sigma$  = standard deviation,  $\mu_1$  = mean of sample 1, and  $\mu_2$  = mean of sample 2.

If you are given standard error (SE) instead of standard deviation, you can calculate standard deviation using the following formula:

$$SE = \frac{SD}{\sqrt{n}}$$

where SE = standard error, SD = standard deviation, and  $n$  = sample size.

The value of  $f(\alpha, \beta)$  in the earlier sample size formula can be obtained using Table 17.3.

**Problem 2** Clinical trials determine efficacy and safety of a drug. Fluoroquinolones (such as ciprofloxacin, levofloxacin, and moxifloxacin) are alternative treatments for drug-resistant *pneumococci* causing community-acquired pneumonia. You wanted to determine the clinical efficacy and safety of two-dose Levaquin (Levofloxacin) regimens in community-acquired pneumonia caused by *pneumococci* bacteria. Your treatment group will receive Levaquin given 750 mg orally in 24 hours for 7 days, and the control group will receive Levaquin given 500 mg orally in 24 hours for 7 days. The treatment will be given to eligible adults in a

**Table 17.3** Use of the data for  $f(\alpha, \beta)$  [5]

		$\beta$			
		0.05	0.1	0.2	0.5
$\alpha$	0.1	10.8	8.6	6.2	2.7
	0.05	13.0	10.5	7.9	3.8
	0.02	15.8	13.0	10.0	5.4
	0.01	17.8	14.9	11.7	6.6

double-blind, controlled clinical trial in a hospital setting. The major outcome variable is days of recovery (assessed by remission of fever, cough, and other symptoms). Calculate the required sample size for a study with 95% confidence and 90% power. Suppose the dropout rate is 10%

### ***Required Information***

From the literature, you found that the recovery using standard doses of Levaquin varies from 10 to 14 days. Data are often mentioned as mean and standard deviation or mean  $\pm$  SD.

$$\text{Mean} \pm \text{SD} = 10 \pm 5.5 \text{ days}$$

$$\text{Mean } 1 (\mu_1) = 10 \text{ days}$$

$$\text{Standard deviation } (\sigma) = 5.5 \text{ days}$$

You also need another important data – Mean 2, the expected mean days of recovery after treatment with the test drug (Levaquin). Suppose it is worthwhile if the patients recover in 7 days.

$$\text{Mean } 2 (\mu_2) = 7 \text{ days}$$

For 95% confidence,  $\alpha = 0.05$ ;

For 90% power,  $\beta = 1 - 0.9 = 0.1$

Using Table 17.3,  $f(\alpha, \beta) = 10.5$

### ***Calculations***

$$n = \frac{2\sigma^2}{(\mu_1 - \mu_2)^2} \times f(\alpha, \beta)$$

$$n = \frac{2(5.5)^2}{(10 - 7)^2} \times 10.5$$

$$n = 71$$

To compensate for the dropout rate, add 10% of 71 = 7.1 to the total.

$$n = 71 + 7 = 78$$

In this clinical trial, you have two groups – one group receiving 750 mg of Levaquin daily and another group receiving the standard dose of 500 mg of Levaquin daily.

The total required sample size =  $2 \times 78 = 156$ .

### 3.3 Sample Size for a Qualitative Study (Using Two Proportions)

Obviously, the formula is different [5]:

$$n = \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_2 - p_1)^2} \times f(\alpha, \beta)$$

Here,  $p_1$  = proportion 1, which is the baseline or control group proportion  
 $p_2$  = proportion 2, which is the expected proportion after the intervention

**Problem 3** Instead of data for mean and standard deviation of the recovery days from treatment, you may have data for the proportion of people recovered within a reasonable treatment period. Our group (Mitra et al. 1995) conducted a double-blind, placebo-controlled clinical trial using hyperimmune bovine colostrum (HBC), which contains colostrum from immunized cows for the treatment of rotavirus diarrhea in children [6]. The study was conducted in 1994 in Bangladesh. Let us assume that you wanted to determine the consistency of the study to be conducted in a different population of a different country. In epidemiology, the consistency of findings substantiates that the results are valid. In your study, you wanted to find out what proportion of children get better in 72 hours (3 days) of treatment between the study drug HBC and a placebo, the colostrum collected from unimmunized cows. Calculate the required sample size for the study, which yields 95% confidence and 90% power. The dropout rate is 15%.

#### Data Required for the Analysis

Suppose,  $p_1$  = proportion of children recovered with placebo in 72 hours = 60% or 0.6  
 $p_2$  = expected proportion of children recovered with the test drug (HBC) in 72 hours = 90% or 0.9

For 95% confidence,  $\alpha = 0.05$ ;

For 90% power,  $\beta = 1 - 0.9 = 0.1$

Using Table 17.3,  $f(\alpha, \beta) = 10.5$

Dropout rate = 15%

#### Calculations

$$n = \frac{p_1(1-p_1) + p_2(1-p_2)}{(p_2 - p_1)^2} \times f(\alpha, \beta)$$

$$n = \frac{0.6(1-0.6) + 0.9(1-0.9)}{(0.6 - 0.9)^2} * 10.5$$

$$n = 39$$

To cover the dropout rate of 15%, add 15% of 39 = 6

$$n = 39 + 6 = 45.$$

The number should be multiplied by 2, because you have two groups.  
The required sample size =  $45 \times 2 = 90$ .

### 3.4 Sample Size for a Case–Control Study Using Odds Ratio (OR)

#### **Required Information [7]**

1. Two of the following should be known:

- $P_1$  = Anticipated probability of exposure for people with the disease

*Formula*

$$P_1 = \frac{a}{a+b}$$

- $P_2$  = Anticipated probability of exposure for people without the disease

*Formula*

$$P_2 = \frac{c}{c+d}$$

- OR = Anticipated odds ratio

$$\text{OR} = \frac{ad}{bc}$$

2. Confidence level = 95%

3. Relative precision =  $\epsilon$

**Problem 4** In a rural district of Bangladesh, a water-borne disease called cholera poses a serious public health problem, especially in winter. About 25% of the people are believed to be using water from contaminated sources such as rivers, canals, and ponds. However, you don't know what proportion of people are exposed to the contaminated water source for people with the disease. Calculate the required sample size for a study with 95% confidence to estimate the odds ratio (OR) to within 25% of the true value (which is the precision). Suppose the true value of OR is 2.5.

**Data**

$P_1$  = anticipated probability of exposure given disease = not known

$P_2$  = anticipated probability of exposure given no disease = 25%

Anticipated OR = 2.5

Confidence = 95%

Relative precision = how close you are from the true value = within 25% of the true value.

There are two ways you can get the solution to the Problem 4. In either case, use the WHO manual [7]. A pdf version of the manual is available at the following link: [https://tbrieder.org/publications/books\\_english/lemeshow\\_samplesize.pdf](https://tbrieder.org/publications/books_english/lemeshow_samplesize.pdf).

- Either use a formula (page 42 of the WHO manual) and calculate the sample size manually, or
- Get the required sample size using a table, which is also available in the WHO manual.

The WHO manual provides tables for several levels of precision and confidence. Just use the correct table. For the data presented in this section in Problem 4, the required sample size = 435, which is available in 6c, page 44 of the WHO manual.

### 3.5 Sample Size for a Cohort Study Using Relative Risk (RR)

#### Required Information [7]

1. Two of the following should be known (using Table 17.4 provided earlier):
  - $P_1$  = Anticipated probability of disease in people exposed to the factor of interest

*Formula*

$$P_1 = \frac{a}{a + c}$$

- $P_2$  = Anticipated probability of disease in people not exposed to the factor of interest

*Formula*

**Table 17.4**  $2 \times 2$  contingency table

Status of disease	Exposure status		Total
	Exposed	Unexposed	
Disease present	$a$	$b$	$a + b$
Disease absent	$c$	$d$	$c + d$
Total	$a + c$	$b + d$	$a + b + c + d$

$$P_2 = \frac{b}{b+d}$$

- RR = Anticipated relative risk

$$\text{RR} = \frac{a}{a+c} / \frac{b}{b+d}$$

2. Confidence level = 95%

3. Relative precision =  $\epsilon$

### **Problem 5**

Suppose you are interested in investigating the link between asbestos exposure and lung cancer using a prospective cohort study design. In the cohort, one group is exposed to asbestos, and another is not. You follow the two groups and collect information on how many get lung cancer in 10 years. From the available data, you know the anticipated probability of lung cancer given no exposure to asbestos, which is 10%.

#### **Data**

$P_1$  = Anticipated probability of disease (lung cancer) in people exposed to the factor of interest (asbestos) = not known

$P_2$  = Anticipated probability of disease (lung cancer) in people not exposed to the factor of interest (no exposure to asbestos) = 10%

RR = 2.25

Confidence level = 95%

Relative precision = 20%, which means that you estimate RR within 20% of the true value in the population.

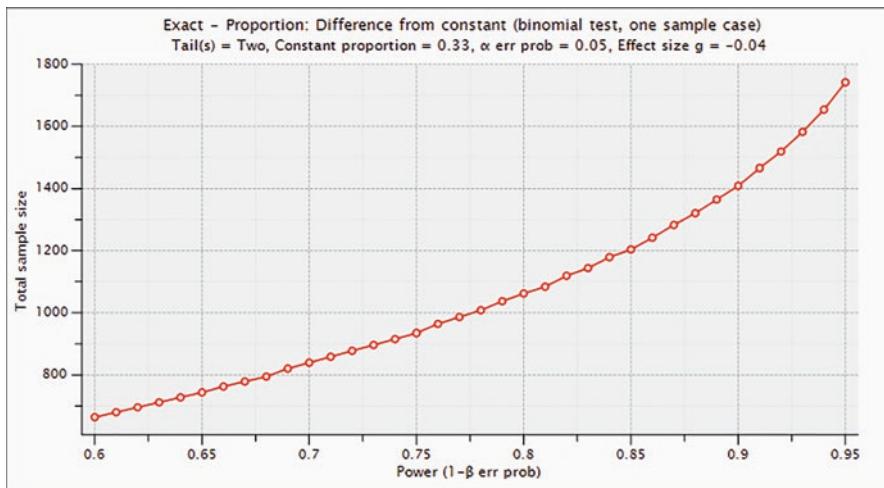
The estimated sample size = 961 (page 53 of the WHO Manual) [7]

The steps of calculating sample size for a case-control study and a prospective cohort study look similar, but the formulae are not the same.

A higher precision value means that the margin of error is higher; it also means you are going further away from the true value in the population. For example, if you use a precision level of 50%, it means you are within 50% of the true value. In that case, the sample size with 95% confidence would be only 100, whereas it was 961 for a precision level of 20%.

### **3.6 Sample Size Estimation Using G\*Power Software**

G\*Power is a free statistical software often used to estimate sample size [8]. Here is a link for downloading the software: <https://gpower.software.informer.com/3.1/>. In this section, I will show a few basic steps in calculating sample size for a cross-sectional study.



**Fig. 17.1** Power curve generated by G\*Power

**Problem 6** You plan to conduct a cross-sectional (prevalence) study to identify the proportion of adolescents with vitamin D deficiency in Malaysia. In a previously conducted study, Quah, et al. (2018) reported that 33% of adolescents were vitamin D deficient [9]. You are going to do a similar study with a precision rate of 4% (meaning your data will be within 4% of the true value), with a 95% confidence ( $\alpha = 0.05$ ), and a power of 80% ( $\beta = 0.2$ ), the dropout rate is 15%. Calculate the required sample size and show the power calculation.

Main Window: Use the steps as follows:

- Use Test Family – select “Exact” (this is for cross-sectional study)
- Statistical test – determine Proportion: Difference from constant.
- Tails – select Two
- Check  $\alpha$  error probability = 0.05 (no change)
- Change Power (your desired number) = 0.80 (the default is 0.95)
- Click Determine (on the left)

A second window pops up. Use the steps as follows:

- Check Difference  $P_2 - P_1$
- $P_1 = 0.33$  (data used from the previous study)
- Your expected difference from the true value = 4%
- $P_2 = 0.29$  (4% change)
- Press Calculate and transfer to the main window. Now recheck the main window.
- Press Calculate (main window) (Don’t make any changes)
- You get the Total sample size = 1062
- Adding the dropout rate of 15%, the total sample size = 1221

For Power Calculation: Use the steps as follows:

- Click on the X-Y plot for a range of values. A third window pops up. Click on Draw plot. You get the Power curve. Save the plot to a file and copy it to your document (Fig. 17.1).

## 4 Further Practice

### A. Problem-Solving Questions

1. Multiresistant *Vibrio cholerae* has been identified in a population. Hospital records show that cholera patients recover in  $5 \pm 2.5$  days (mean  $\pm$  SD) following treatment with erythromycin. Suppose you want to investigate if another drug, ciprofloxacin, could cure the patients earlier, with anticipated recovery days of 3.5. Calculate the required sample size for a clinical trial having 95% confidence and 80% power. Note that 5% of the patients may be dropped after they have been enrolled in the study. Answer the following questions using the Problem in Q1.
  - (a) Which formula should you use in calculating the sample size for the data provided in question no. 1?
  - (b) What is the constant value of  $\alpha, \beta$  for 95% confidence and 80% power (use table)?
  - (c) How many groups for the study?
  - (d) What is the dropout rate? What should you do with the sample size for the dropout?
  - (e) What is the required total sample size?
2. In Kelantan, Malaysia, the stunting rate among young children is 34%. You hypothesize that using complementary milk feeding with essential micro-nutrients for six months, the stunting rate will decrease by 10%. Calculate the required sample size for a study with 95% confidence and 90% power. The dropout rate is 20%.
3. Power and sample size are inversely proportional True/False
4. Interpret the statement - “ $p < 0.0001$ , the data are highly significant.”
5.  $\alpha = 0.01$ , calculate a confidence
6.  $\beta = 0.1$ , calculate power
7. What is the relationship between effect size and sample size?
8. What do you mean by relative precision? Use an example to illustrate it.
9. What do you mean by Type I error?
10. When do you use a two-tailed test? Why is it more common to use a two-tailed test?

## Answer Keys

1. a. Quantitative formula because you have data for mean and standard deviation.  
b. For 95% confidence,  $\alpha = 0.05$ ; for 80% power,  $\beta = 0.2$ ; the value of the constant (from Table 17.3) = 7.9.  
c. Two groups – multiply the calculated sample size by 2 for the total sample size.  
d. The dropout rate is 5%; add 5% of the calculated sample size to get the required total sample size for the study.  
e. The required total sample size is 92.
2. Use the qualitative formula of proportion.  $P_1 = 34\% = 0.34$ ;  $P_2 = 24\%$  (10 percent reduction) = 0.24, with  $\alpha = 0.05$  and  $\beta = 0.2$ , the factor of  $\alpha$  and  $\beta = 10.5$ . The calculated sample size = 427 for each group;  $427 \times 2 = 854$  total; with 20% dropout, the total sample size = 1026.
3. False;
4. If the  $p$ -value is  $\leq 0.05$ , the data between the groups are statistically significant. It is suggested not to mention that data “are highly significant” or “almost significant.” Just mention it is significant or not.
5. Confidence =  $1 - \alpha$ ;  $1 - 0.01 = 0.99$  or 99%.
6. Power =  $1 - \beta$ ;  $1 - 0.1 = 0.9$  or 90%.
7. If the effect size is increased, the sample size is decreased. They vary inversely.
8. Relative precision means how close you are from the true value. The relative precision of 25% means that the estimate (either odds ratio for a case-control study or RR for a cohort study) is within 25% of the true value.
9. Type I error means rejecting a true null hypothesis.
10. When the outcome is uncertain (it can go in either cure or not cure), we use a two-tailed test. It is most commonly used because we do not know the outcome positive or negative in most situations.

## References

1. File TM Jr. Clinical implications and treatment of multiresistant *Streptococcus pneumoniae* pneumonia. Clin Microbiol Infect. 2006;12(Suppl. 3):31–41.
2. Cowling T, Farrah K. Fluoroquinolones for the treatment of respiratory tract infections: a review of clinical effectiveness, cost-effectiveness, and guidelines. Canadian Agency for Drugs and Technologies in Health; 2019. Available at: <https://www.ncbi.nlm.nih.gov/books/NBK543514/>.
3. Daniel WW, Cross CL. Biostatistics: a foundation for analysis in the health sciences. 10th ed. Hoboken: Wiley; 2013.
4. Cardoso A, Silva A, Marín H. Pregnant women's knowledge gaps about breastfeeding in Northern Portugal. Open J Obstet Gynecol. 2017;7:376–85. <https://doi.org/10.4236/ojog.2017.73039>.
5. Pocock SJ. Clinical trials – a practical approach. 1st ed. New York: Wiley; 1983.
6. Mitra AK, Mahalanabis D, Ashraf H, Unicomb L, Eeckels R, Tzipori S. Hyperimmune cow colostrum reduces diarrhoea due to rotavirus: a double-blind, controlled clinical trial. Acta Paediatr. 1995;84:996–1001.

7. Lwanga SK, Lemeshow S. Sample size determination in health studies. A practical manual. Geneva: World Health Organization; 1991. [https://tbrieder.org/publications/books\\_english/lemeshow\\_samplesize.pdf](https://tbrieder.org/publications/books_english/lemeshow_samplesize.pdf)
8. Kang H. Sample size determination and power analysis using the G\*Power software. J Educ Eval Health Prof. 2021;18:17. <https://doi.org/10.3352/jeehp.2021.18.17>. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8441096/pdf/jeehp-18-17.pdf>
9. Quah SW, Abdul Majid H, Al-Sadat N, Yahya A, Su TT, Jalaludin MY. Risk factors of vitamin D deficiency among 15-year-old adolescents participating in the Malaysian Health and Adolescents Longitudinal Research Team Study (MyHeARTs). PLoS One. 2018;13(7):e0200736. <https://doi.org/10.1371/journal.pone.0200736>.

# Chapter 18

## Missing Data Imputation: A Practical Guide



Enayetur Raheem

### Learning Objectives

After completing this chapter, you will be able to:

- Understand the issue at hand through a real data set involving missing observations
- Understand different classes of missing data and missing data mechanisms
- Identify the types of missingness in the data
- Learn the commonly used strategies for handling missing data and how to apply these strategies
- Apply multiple imputations and parameter estimation from the imputed data sets

## 1 Introduction

Missing data is inevitable in research. Most of the researchers face issues with missing values when analyzing data. This chapter presents missing data problems and discusses how to handle them. The methods are briefly explained from a practitioner's point of view and therefore avoided theory or mathematical formulations altogether. A real data set containing missing values on multiple variables was used to demonstrate the use of various missing value handling approaches.

From a practical perspective, researchers encounter the issue of missing value when a bivariate analysis is performed. In the presence of a missing value, the relationship between the likely factor or factors and the primary outcome of interest is often unclear. Different approaches for handling missing values have been discussed and demonstrated in the following. Since the handling of missing values must be determined at the start of the analysis, the researcher needs to understand the pattern and types of missing values in the data. Unfortunately, there is no single recipe that would work for all situations. Instead, it is situation-specific and based on the contexts that may have caused the missingness.

---

E. Raheem (✉)  
Biomedical Research Foundation, Dhaka, Bangladesh  
e-mail: [enayetur.raheem@brfbd.org](mailto:enayetur.raheem@brfbd.org)

If there are only a few missing values in a negligible number of records, then listwise deletion, discussed later in this chapter, is likely a simple fix. However, imputation of some sort would be necessary if missingness is not negligible. Among the many techniques discussed, the least favored are those involving imputation by mean, median, or mode. The recommended ones are those based on hot-deck, KNN, and multiple imputations. Among them, multiple imputations are the most widely used in various fields.

Although multiple imputations are preferred in many missing data problems, there may be situations where alternative methods are just fine. One example is weighting-based methods and likelihood-based methods. Likelihood-based methods which involve Expectation–Maximization (EM) algorithm are in some sense known as the “royal way” to handle missing values [1]. For further reading on these likelihood-based methods, one can find an excellent book on EM algorithm and related methods by Little and Rubin (2019) [2]. This chapter provides guidelines and strategies for dealing with missing values in data analysis.

## 2 Real-Life Example

### 2.1 *Hepatocellular Carcinoma Data*

Let us consider a typical scenario when a statistician, epidemiologist, or anyone doing the analysis creates the descriptive summary table, as shown in Table 18.1. Such a table is also popularly known as “Table One.” Consider the Hepatocellular Carcinoma study data collected at a University Hospital in Portugal. The data consists of clinical records of 165 patients diagnosed with HCC. The data was downloaded from the UCI Machine Learning Repository at <https://archive-beta.ics.uci.edu/>. There are 49 features and 1 outcome/target variable. The target variable was encoded as a binary variable with 0 and 1, 1 indicating survival at one year and 0 indicating otherwise. For details on the clinical features and the analysis performed, please see Santos et al. [3].

To orient the readers to the data set, let us briefly describe the selected variables and their types in the following. It will be important later when one performs statistical tests and other missing value analyses.

It is known from the medical literature that the common factors for HCC include chronic hepatitis (such as hepatitis B and hepatitis C infections) and cirrhosis of the liver. Cirrhosis of the liver is assessed from five clinical measures of the liver, namely, total bilirubin, albumin, encephalopathy, ascites, and prothrombin time INR. Cirrhosis of the liver is present in over 80% of HCC cases, clearly identified as this pathology’s main precursor lesion. One would select a subset of the risk factors to demonstrate how missing data appears in statistical analysis. Next, statistical analysis is performed. Note that the analysis performed here is not intended to demonstrate how the actual HCC data would be analyzed. The following is a

**Table 18.1** Variable description

Variable	Description	Type
gender	Gender	Binary (1 = Male, 0 = Female)
symptoms	Symptoms	Binary (1 = Present, 0 = Absent)
hbsag	Hep B-positive	Binary (1 = Positive, 0 = Negative)
hcavab	Hep C-positive	Binary (1 = Present, 0 = Absent)
cirrhosis	Cirrhosis	Binary (1 = Present, 0 = Absent)
smoking	Smoker	Binary (1 = Present, 0 = Absent)
diabetes	Diabetic	Binary (1 = Present, 0 = Absent)
hemochromatosis	Hemochromatosis	Binary (1 = Present, 0 = Absent)
esophageal_varices	Esophageal varices	Binary (1 = Present, 0 = Absent)
age_at_dx	Age at diagnosis	Quantitative/ratio
performance_status	Performance status	Ordinal
creatinine	Creatinine	Quantitative/ratio
hemoglobin	Hemoglobin	Quantitative/ratio
iron	Iron	Quantitative/ratio
ferritin	Ferritin	Quantitative/ratio

demonstration to highlight the missing value issue and what approaches can be taken to handle them.

Table 18.2 shows a bivariate analysis of the selected risk factors and the one-year survival of HCC patients. The rows represent the risk factors, while the column represents two categories of the outcome variable – the patient survived or did not survive. The *p*-value column shows the *p*-value associated with either Pearson's Chi-squared test, Fisher's exact test, or Wilcoxon rank sum test. For categorical variables, *n* (%) was reported, while the median (and interquartile range) was reported for ratio scale variables.

It is noted that quite a sizable number of missing values are present in the data. For example, 23 deceased patients did not have data on ferritin, while 56 surviving patients did not. Complete data are available on gender, cirrhosis, and age at diagnosis. Also, there is no missing value in the outcome variable of interest.

The bivariate results show that some risk factors significantly affect the one-year survival while some do not. However, since missing values exist in the data, it is essential to understand whether the significance is true because of the risk factor alone or whether the missing observations have any bearing.

The table presented here was created using the “*gtsummary*” library in R. Depending on what software is used, the missing values are handled and reported differently based on the default settings in these statistical packages. While descriptive summaries such as count and percentage would either report missing values or ignore them, statistical software exclusively removes missing values, often without warning, when calculating mean, median, and other measures of location. Also, when performing statistical models, all software will remove all rows containing missing values in any of the columns (risk factors). This is known as *listwise*

**Table 18.2** Bivariate analysis of risk factors and one-year survival status

Characteristic	Deceased, N = 63	Survived, N = 101	p-Value
Gender (Male)	52 (83%)	80 (79%)	0.6
Any symptoms	50 (81%)	44 (52%)	<0.001
(Missing)	1	17	
Hep B-positive	5 (9.4%)	11 (12%)	0.7
(Missing)	10	7	
Hep C-positive	16 (28%)	18 (18%)	0.2
(Missing)	6	3	
Cirrhosis	56 (89%)	92 (91%)	0.6
Smoker	21 (46%)	41 (53%)	0.4
(Missing)	17	24	
Diabetic	26 (41%)	29 (30%)	0.13
(Missing)	0	3	
Hemochromatosis	3 (6.4%)	3 (3.2%)	0.4
(Missing)	16	7	
Esophageal varices	20 (57%)	48 (62%)	0.6
(Missing)	28	24	
Age at diagnosis	69 (60, 78)	64 (56, 73)	0.035
Performance status			<0.001
0	17 (27%)	62 (61%)	
1	14 (22%)	16 (16%)	
2	15 (24%)	17 (17%)	
3	12 (19%)	6 (5.9%)	
4	5 (7.9%)	0 (0%)	
Hemoglobin	12.20 (10.80, 13.20)	13.70 (12.03, 14.90)	<0.001
(Missing)	0	3	
Total bilirubin	1.70 (1.00, 3.70)	1.30 (0.80, 2.42)	0.031
(Missing)	0	5	
Creatinine	0.90 (0.74, 1.29)	0.80 (0.70, 1.03)	0.11
(Missing)	2	5	
Iron	52 (37, 102)	92 (57, 144)	0.009
(Missing)	24	54	
Ferritin	538 (175, 838)	239 (77, 363)	0.002
(Missing)	23	56	

*deletion.* Such an approach is not feasible for this data as there are 29 complete cases out of 164 records in our working data set.

Thus, let us first explore the data to understand the completeness of records.

### 3 Exploring Missingness

In an ideal scenario, there would be no missing data in any variables. However, that is rarely the case. Missing data is inevitable. As such, handling missing data is essential before doing the actual analysis. Thus, missing data should be explored at the exploratory data analysis (EDA) step. However, although EDA is a routine step in all analytic situations, thorough exploration of missing values is rarely performed. This is mainly because most analysis is performed after listwise deletion. However, that comes with a price of loss of statistical power, especially if the number of records is small to begin with.

Many statistical packages have features to explore missing data. Some popular libraries that can explore missing data in R include *skimr*, *naniar*, *mice*, and *VIM*. First, we want to know the percentage of records that are complete. We use the *naniar package* to explore it visually. In the following demonstration, the data set is stored as an R data frame object named hcc.

```
library(naniar)  
gg_miss_var(hcc, show_pct = TRUE)
```

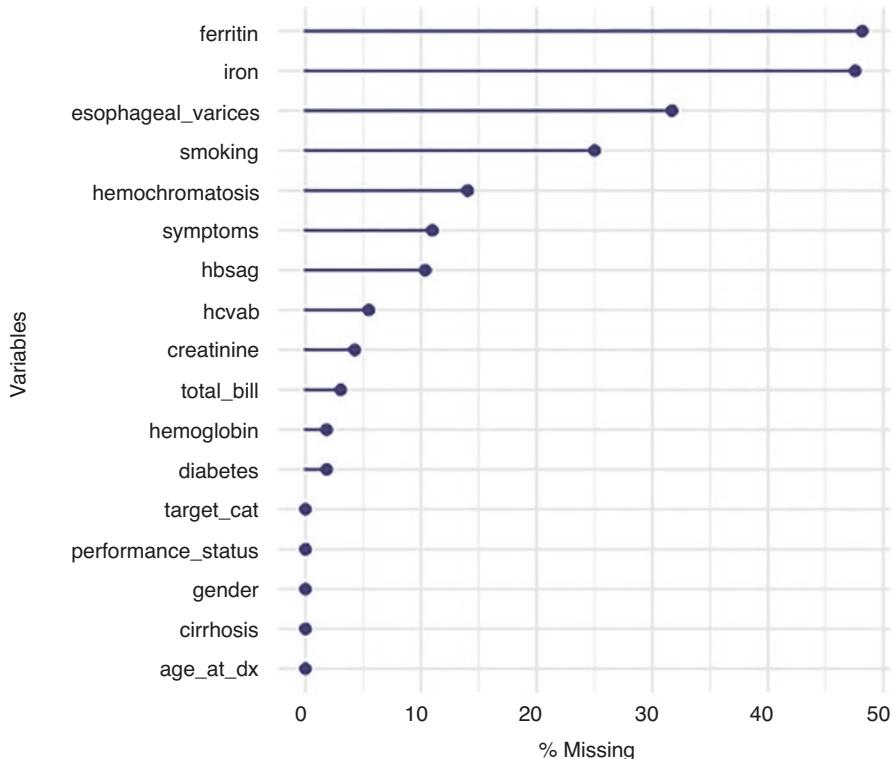
We first loaded the *naniar* library using the `library(naniar)` command in the above commands. Then we used `gg_miss_var()` function to create Fig. 18.1. The `show_pct = TRUE` indicates that we want the percentage of the missing values displayed in the figure. Figure 18.1 shows the percentage of missing observations (on the x-axis) and the variable names plotted on the y-axis.

## 4 Classification of Missing Data

The classification of missing data and related terminologies was introduced by Rubin [4] and his colleagues. Based on their work, three mechanisms for missing data have been developed. These are missing at random (MAR), missing completely at random (MCAR), and missing not at random (MNAR).

### 4.1 Missing Completely at Random (MCAR)

This type of missing data arises when the missingness is truly random. That is, the observed data has no relationship with the missing data. In other words, missing is completely random when the probability of the missing value of a variable  $Y$  is completely unrelated to the missing value in another variable in the data. Also, the missingness is not related to the variable  $Y$  itself. Put another way, when the chance of missing data is the same for all the variables, it is MCAR. For example, in a



**Fig. 18.1** Percentage of missing values by each variable

sample survey, if the data are collected following a statistical design, every element in the population has an equal chance of being selected. The elements not selected in the sample can be said to be MCAR.

As it turns out, MCAR is the strictest form of missingness and is somewhat unrealistic, as missing data occurs in statistical sampling, too.

## 4.2 Missing at Random (MAR)

This is a more realistic and practical missing mechanism. Data are missing at random (MAR) when the missingness is dependent on other variable or variables within the observed data but unrelated to the variable  $Y$  itself. This definition could have been more explicit at first. A more straightforward way to understand it is “if the probability of being missing is the same only within groups defined by the observed data” [1], then the data is considered MAR. From a statistical perspective, missing data is random if the missingness in  $Y$  is unrelated to  $Y$  after controlling for the rest of the variables in the data.

In contrast with MCAR, where the missingness is unrelated to anything in the observed data, MAR only requires that the missingness is not due to the variable itself. However, it may be due to some other variable within the data.

MCAR is thus a stricter version of MAR, making the MAR a much broader class in the missing data mechanism.

### 4.3 Missing Not at Random (MNAR)

Data on  $Y$  is considered missing, not at random, if the missingness is dependent on the variable itself after adjusting for the other variables. For example, missing value on the patient's performance score could be due to the patients being withdrawn from the study due to poor health issues. In this case, the missingness directly relates to what is being measured.

## 5 Identifying Missing Data Mechanism

Now that we know the missing data mechanism, the critical question is whether we can identify the type of missingness in our data.

### 5.1 Testing MCAR

It is possible to test for MCAR since it assumes that the observed data are a random subset of the hypothetically *complete* population which the samples were selected from. To test if the data are MCAR, consider the data we presented earlier. About 35% of records are missing on esophageal varices. If the data are MCAR, then there should not be a significant difference between the observed and missing records between the one-year survival group. To test this, we consider the missing value a valid category when performing the Chi-squared test. Table 18.3 shows a statistical test for esophageal varices before and after factoring in the missing values. We cannot claim MCAR assumption since there were significant differences in esophageal varices when missing values were considered a valid category.

**Table 18.3** Testing MCAR for esophageal varices by one-year survival

Characteristic	Deceased, $N = 63$	Survived, $N = 101$	$p$ -Value
Esophageal varices	20 (57%)	48 (62%)	0.6
(Missing)	28	24	
Esophageal varices (MCAR test)	35 (56%)	77 (76%)	0.006

```
# Load library
library(tidyverse)

hcc %>%
  mutate(
    esophageal_varices_mcar =
      case_when(is.na(esophageal_varices) ~ 0, TRUE ~ 1)
  ) %>%
  select(
    esophageal_varices, esophageal_varices_mcar, target_cat
  ) %>%
 tbl_summary(
  by = target_cat,
  missing_text = '(Missing)',
  label = list(
    esophageal_varices ~ 'Esophageal varices',
    esophageal_varices_mcar ~ 'Esophageal varices (MCAR test)'
  )
) %>%
add_p() %>%
modify_caption("Testing MCAR for Esophageal varices
by 1-year survival.")
```

We could do a similar test for a ratio scale variable too. For that, we would create binary variables indicating missing data and then perform a t-test for the means between the groups (complete records vs. incomplete records). Let us implement it below for predicting hemoglobin status using ferritin.

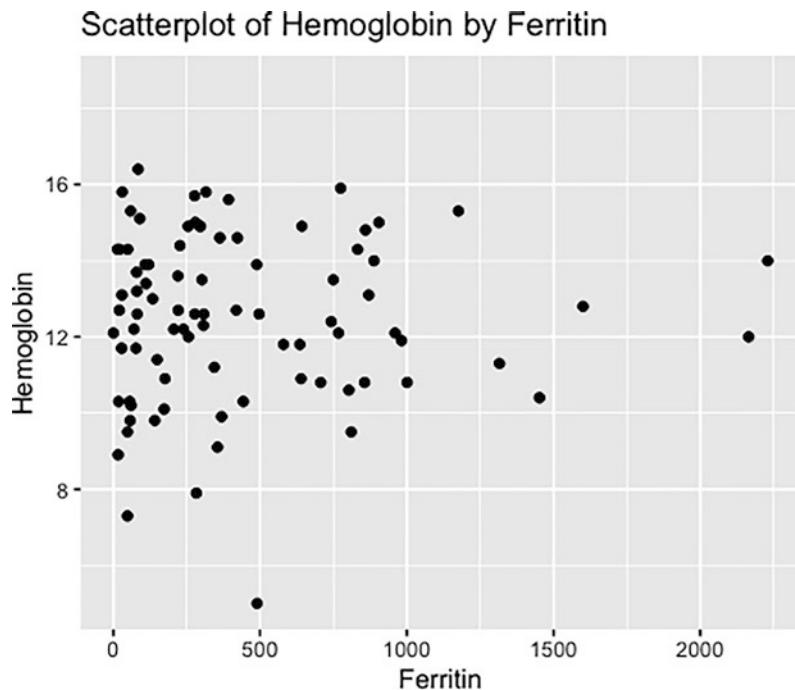
```
hcc %>%
  ggplot() +
  aes(x = ferritin, y = hemoglobin) +
  geom_point() +
  labs(title="Scatterplot of Hemoglobin by Ferritin",
       x ="Ferritin", y ="Hemoglobin")

## Warning: Removed 79 rows containing missing values (geom_point).
```

Notice in Fig. 18.2 that `ggplot()` printed a warning message indicating the removal of missing values before drawing the scatterplot. To better assess the relationship visually, we want to highlight the missing data points on the scatterplot. We redraw the scatterplot by survival status.

```
hcc %>%
  ggplot(aes(x = ferritin, y = hemoglobin)) +
  geom_miss_point() +
  facet_wrap(~ target_cat) +
  theme(legend.position="top")
```

In Fig. 18.3, the red dots along the y-axis represent hemoglobin values when the ferritin data are missing. There is a noticeable missingness in ferritin for the patient



**Fig. 18.2** Scatterplot of hemoglobin and ferritin

who survived compared to those deceased. Let us perform the statistical test to confirm if the difference is significant. The results are shown in Table 18.4.

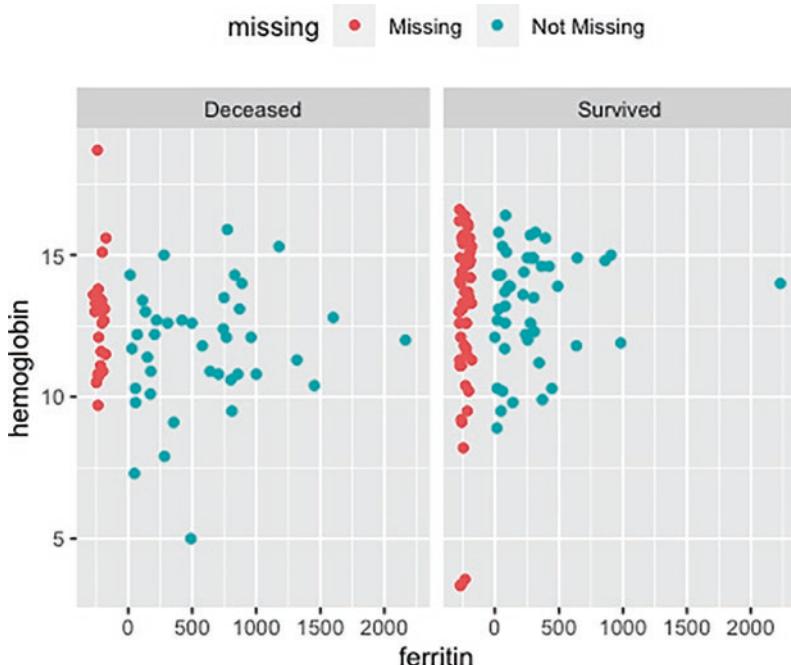
```

hcc %>%
  mutate(
    ferritin_bin = case_when(
      is.na(ferritin) ~ 0, TRUE ~ 1
    )
  ) %>%
  select(hemoglobin, ferritin_bin) %>%
 tbl_summary(by = ferritin_bin, missing = 'no')
) %>%
add_p(test = list(hemoglobin ~ 'wilcox.test'))

```

In the gtsummary library, the Wilcox test is used by default for a continuous variable when the by variable has two levels. One can perform a t-test instead. To do so, modify the option as `add_p(test = list(hemoglobin ~ 't.test'))`. The test procedure did not change the decision, however. When the missing values were accounted for, we found a somewhat significant difference in hemoglobin between the survival groups. This may suggest that the missing data mechanism is not MCAR.

Comparing the density or histogram by group makes more sense for a continuous variable. See Fig. 18.4 for the density plot of hemoglobin by survival status. Notice



**Fig. 18.3** Scatterplot of hemoglobin against ferritin by survival status

**Table 18.4** Testing for statistical difference of hemoglobin in the two outcome groups

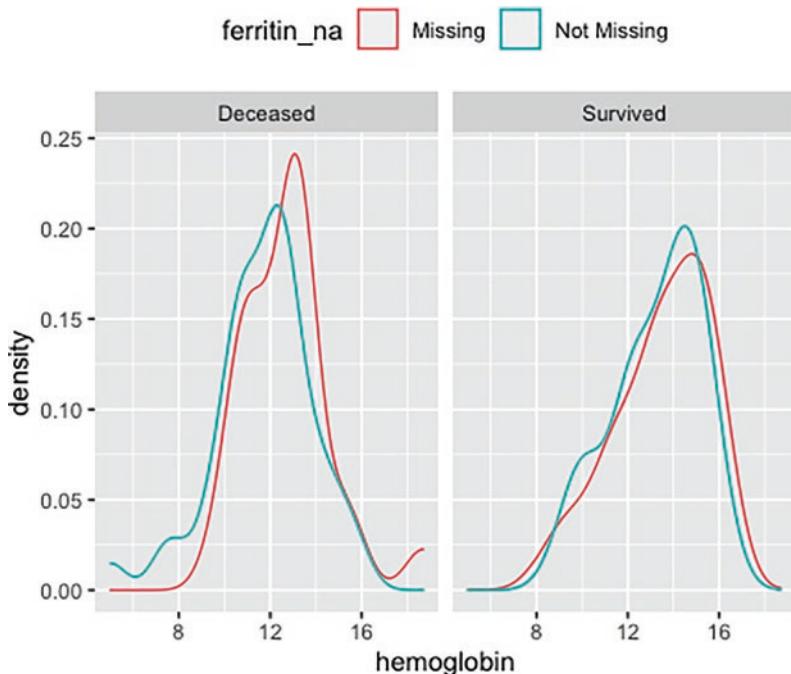
Characteristic	0, N = 79	1, N = 85	p-Value
Hemoglobin	13.30 (11.67, 14.90)	12.60 (10.90, 14.30)	0.046

that three records were removed because the hemoglobin variable was missing on those records.

```

hcc %>%
  mutate(
    ferritin_na = case_when(
      is.na(ferritin) ~ "Missing", TRUE ~ "Not Missing"
    )
  ) %>%
  ggplot(aes(x = hemoglobin, color = ferritin_na)) +
  geom_density() +
  facet_wrap(~ target_cat) +
  theme(legend.position="top")

## Warning: Removed 3 rows containing non-finite values (stat_density).
  
```



**Fig. 18.4** Density plot of a continuous variable (hemoglobin) by survival status

## 5.2 Testing MAR and MNAR

We have demonstrated how to test MCAR assumption using a statistical approach. However, MAR, by definition, depends on other variables. Thus, we need to know the missing values to test this mechanism.

Similarly, MNAR assumption is also impossible to validate because missingness depends on the variable itself, and we need to observe the missing values to test the dependence.

This poses a challenge as most statistical inference and missing value imputation techniques rely on the MAR assumption. However, some suggestions in the literature exist to use proxy variables to account for the missingness.

## 6 Common Methods for Missing Data Handling

There are several ways to work with data when missing values are present. In the following, we illustrate how these approaches work in practice. Frequently used approaches include removing the records (rows of the data) if there is a missing value in any variables. These are often known as ad-hoc methods as they do not

attempt to impute or replace the missing values but instead remove them to complete the available data. On the other hand, some methods impute the missing value instead. After evaluating missingness in our data, our next best step is to impute the missing values. Imputation can be performed on a single variable and multiple variables.

## 6.1 Listwise Deletion

This default setup handles missing values in all major statistical packages, including SAS, SPSS, R, and Stata. Listwise deletion is also known as complete case analysis, in which all rows are removed if missing values exist in one or more variables. To understand this, imagine we have ten rows and two variables in the data set. If we have two missing in the first variable on the first and second row and one missing in the third row of the data, then the listwise deletion will remove the data's first, second, and third rows.

If the data are missing completely at random (MCAR), then listwise deletion does not affect the estimated mean, variance, and regression weights. However, if the data are not MCAR, the procedure will severely bias these estimates.

In the HCC data set introduced earlier, we have 164 observations and 17 variables. As noted in Table 18.1, many of the columns in the data have missing values. A listwise deletion can be performed using `na.omit()` function in base R, or `drop_na()` function in the tidyverse package in R.

```
hcc %>% dim()
## [1] 164 17

# Loading the library
library(tidyverse)

# demonstration of listwise deletion and the effect
# on the number of complete records
hcc %>% drop_na() %>% nrow()

## [1] 29
```

If we wish to keep all 17 variables for our analysis, listwise deletion severely reduces the number of complete records.

Another potential issue with listwise deletion is that there will be inconsistencies in the analysis regarding how many complete cases are being utilized. For example, if we only use gender, symptoms, iron, ferritin, age\_at\_dx, and smoking in our analysis, then the number of complete records is much different, as shown below.

```
hcc %>%
  select(gender, symptoms, iron, ferritin, age_at_dx) %>%
  drop_na() %>% nrow()

## [1] 73
```

Clearly, the effect of listwise deletion depends on missingness in each variable and what variables are in the analysis.

## 6.2 Pairwise Deletion

Pairwise deletion is applicable for linear regression and factor analysis-related situations. Pairwise deletion is also known as available case analysis. This is different from listwise deletion, which does not remove all the rows if there is missing values in any of the columns. Instead, the means and covariances are calculated based on the available data in pairwise deletion.

To understand this, we have three variables, and one variable has missing values in some of the rows. Rather than removing all the rows with the missing values, the statistical procedure can still use two remaining variables with complete records. This is advantageous since all the available data are being utilized. However, a disadvantage of this method is that different statistical procedures may use a different subset of the data based on data availability. This may or may not be problematic, depending on a case-by-case basis.

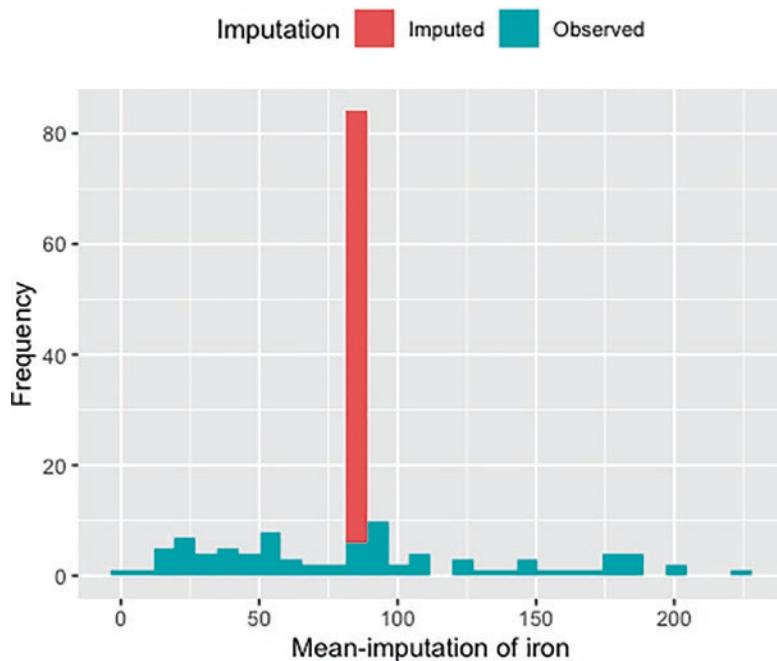
Like the listwise deletion, if the data are not MCAR, then the estimates will be biased if pairwise deletion is applied. This method is not commonly recommended because of the complexity of the estimation and interpretation of results. Statistical software packages commonly have options to calculate covariance and means using pairwise deletion, but incorporating these into the regression model needs special libraries. For example, R's lavaan package can utilize the estimated mean and covariance obtained using pairwise deletion. For a demonstration of the technique using R, please find the study of Van Buuren [1].

## 6.3 Mean or Median Imputation

The most commonly used fix for missing data is replacing them with the variable's mean or median. This applies to numerical variables only. For categorical variables, we may use mode instead. Although this is simple and quick to retain the records with missing values, it may underestimate the standard error of the estimates. Additionally, mean or imputation does not preserve the relationship between variables. Mean or median imputation may severely affect the data distribution. These are recommended for a quick fix when only a few missing values exist.

```
library(simputation)
hcc %>%
  mutate(
    imp_flag = ifelse(is.na(iron), 'Imputed', 'Observed'),
    iron_imp = impute_mean(iron)
  ) %>%
ggplot(aes(x = iron_imp, fill = imp_flag)) +
  geom_histogram() +
  labs(
    x = 'Mean-imputation of iron',
    y = 'Frequency',
    fill = 'Imputation'
  ) +
  theme(legend.position = "top")
```

Using the mean, we imputed the iron variable and demonstrated how the distribution is affected. Figure 18.5 visually demonstrates it. Mean or median imputation is generally only recommended for a quick fix when a few missing values exist.



**Fig. 18.5** Distribution of iron after mean imputation

## 6.4 Regression Imputation

A better way to impute a single variable is to use information from other variables. This is achieved in several ways, and regression is one of them. In this approach, we can build a model using available data and the fitted model to predict missing values and use them to substitute missing values. The researcher should decide which variables to use for regression-based imputation. In the following, we demonstrate using HCC data.

```

hcc2 <- hcc %>%
  mutate(
    imp_flag = if_else(is.na(iron), 'Imputed', 'Observed')
  ) %>%
  impute_lm(iron ~ age_at_dx + total_bill)

hcc2 %>%
  ggplot(aes(x = iron, fill = imp_flag)) +
  geom_histogram() +
  labs(
    x = 'Regression Imputation of iron',
    y = 'Frequency',
    fill = 'Imputation'
  ) +
  theme(legend.position="top")

## Warning: Removed 4 rows containing non-finite values (stat_bin).

```

Here, we are imputing iron using the age\_at\_dx and the total bilirubin. Figure 18.6 shows the distribution after imputation. Note that the distribution is much better

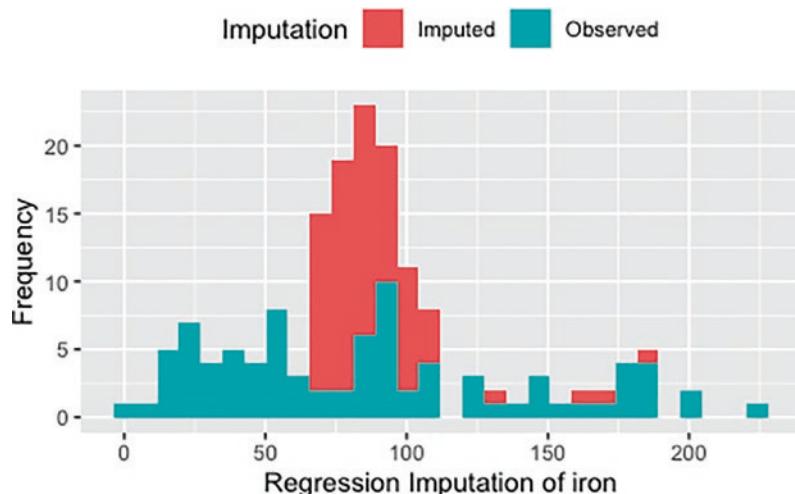


Fig. 18.6 Distribution of iron after regression imputation

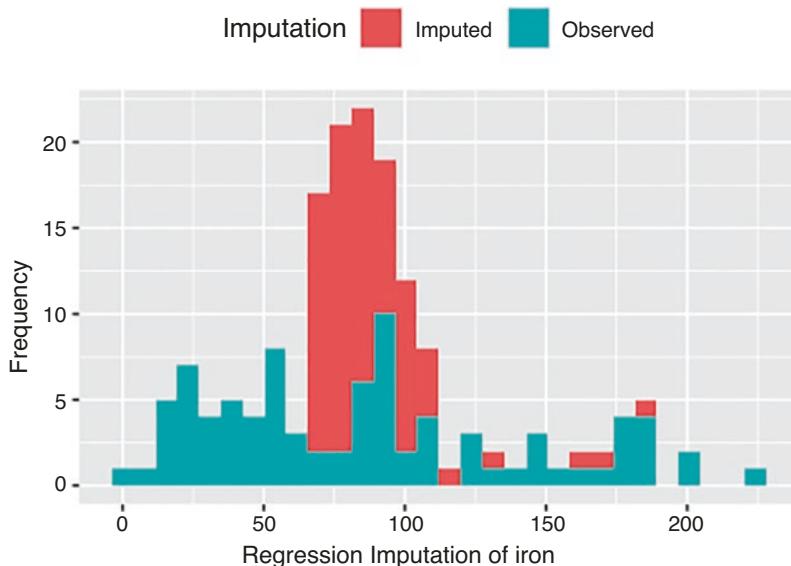
than the one obtained using mean imputation. However, notice here that there are still some missing values in the total\_bill variable, which can be improved by first imputing them using mean imputation. We can perform a chain of imputation to achieve this as follows. The distribution after imputation is displayed in Fig. 18.7.

```

hcc2 <- hcc %>%
  mutate(
    imp_flag = if_else(is.na(iron), 'Imputed', 'Observed'),
    # Fill the missing values in total_bill
    # using mean imputation
    total_bill_imp = impute_mean(total_bill)
  ) %>%
  # chain the regression imputation using the
  # imputed total_bill column
  impute_lm(iron ~ age_at_dx + total_bill_imp)

# Plot the data to visualize the distribution
hcc2 %>%
  ggplot(aes(x = iron, fill = imp_flag)) +
  geom_histogram() +
  labs(
    x = 'Regression Imputation of iron',
    y = 'Frequency',
    fill = 'Imputation'
  ) + theme(legend.position="top")

```



**Fig. 18.7** Distribution of iron after mean imputation followed by a regression imputation

## 6.5 Hot-Deck Imputation

An intuitive imputation method is based on borrowing values from a related donor to substitute the missing value. At the core, a record with a missing value receives a value to fill the missing point from a randomly chosen record that is maximally similar to the record having the missing value. That is, several matching donor records are identified first based on the criteria set by the researcher. The collection of donor cases is called the “deck.” “Historically, the term ‘hot deck’ comes from the use of computer punch cards for data storage and refers to the deck of cards for donors available for a non-respondent [5].” The deck is “hot” because that is the deck actively being used. A “cold deck,” on the other hand, indicates data that are already collected or otherwise processed.

One important advantage of hot-deck imputation is that it is reasonably suitable for situations where the missingness is at random (MAR). Hot deck is most popular for imputing survey data as the data are collected using a structured questionnaire with a defined purpose. Because of the random borrowing of values from similar cases, the marginal distribution of the imputed data is expected to remain unchanged. However, this technique may affect correlations and produce biased regression coefficients.

The idea of hot deck can be used for more complex imputation in several ways, for example, borrowing donors with or without replacement. In without replacement scenario, the donor can only donate once. Without replacement scenarios, the donor record can be used to imputing more than once. Hot-deck imputation can also take only partially matched donors from among the list of matching criteria.

Several hot-deck methods have been implemented in statistical packages. In the simputation R package, three kinds of hot-deck implementation are available: random hot-deck imputation, sequential hot-deck imputation, predictive mean matching, and K-nearest neighbor imputation. We briefly discuss them below.

### 6.5.1 Random Hot-Deck Imputation

The missing value is substituted in this method from a randomly selected donor record. This can be applied to numeric, categorical, or mixed data. To impute all the missing values in the data using random hot-deck imputation, use the `impute_rhd()` function as shown below:

```
hcc_rhd <- impute_rhd(hcc, . ~ target_cat, pool = 'complete')
```

The `pool= 'complete'` indicates that the donor pool consists of records with complete observations on the left-hand side of the model. In the example above, all but

the target\_cat variable is being used to indicate the donor pool. There are other options, namely univariate and multivariate. For details, please see the documentation of the simputation package [6].

### 6.5.2 Sequential Hot-Deck Imputation

This approach uses either the last observation carried forward (LOCF) or the next observation carried backward (NOCB). First, the data set was sorted using the variables used as predictors. Missing values are substituted using LOCF or NOCB as required. The LOCF imputation is a simple idea where the missing value is copied from the immediate past record. The NOCB works similarly, except the missing value is replaced using a value from the next record. These methods are most common in longitudinal studies.

```
hcc_shd <- impute_shd(hcc, . ~ target_cat, pool = 'complete',
                        order = 'nocb')
```

The function `impute_shd()` is similar to the previous one with an added parameter to set the order – either `locf` or `nocb`. This method applies to both categorical and numeric data.

### 6.5.3 Predicting Mean Matching Imputation

As the name suggests, predictive mean matching predicts the missing value using a regression model. The predicted value is not used to substitute the missing value; rather, a combination of values nearest to the predicted value is used to impute the missing value. Thus, this method can only be applied to numerical data. The nearest value is the non-missing value with the smallest absolute deviation from the prediction. The `impute_pmm()` function in `simputation` package can do it.

## 6.6 K-Nearest Neighbor Imputation

There are several machine learning-based algorithms to impute missing values. The commonly used one is based on the K-nearest neighbor. This method aims to identify similar complete records based on some criteria, and then a suitable donor is selected from among them. In the `impute_knn()` function, the similarity is determined based on the similarity coefficient described by Gower (1971) [7].

**Table 18.5** Bivariate analysis of risk factors and one-year survival status on the imputed data after KNN imputation

Characteristic	Deceased, N = 63	Survived, N = 101	p-Value
Gender(male)	52 (83%)	80 (79%)	0.6
Any symptoms	49 (78%)	54 (53%)	0.002
Hep B-positive	7 (11%)	12 (12%)	0.9
Hep C-positive	16 (25%)	20 (20%)	0.4
Cirrhosis	56 (89%)	92 (91%)	0.6
Smoker	33 (52%)	51 (50%)	0.8
Diabetic	27 (43%)	28 (28%)	0.046
Hemochromatosis	4 (6.3%)	5 (5.0%)	0.7
Esophageal varices	31 (49%)	56 (55%)	0.4
Age at diagnosis	68 (60, 76)	64 (57, 73)	0.15
Performance status			<0.001
0	17 (27%)	62 (61%)	
1	15 (24%)	15 (15%)	
2	15 (24%)	17 (17%)	
3	11 (17%)	7 (6.9%)	
4	5 (7.9%)	0 (0%)	
Hemoglobin	12.20 (10.85, 13.20)	13.70 (12.00, 14.90)	<0.001
Total bilirubin	1.90 (1.00, 3.85)	1.30 (0.80, 2.40)	0.008
Creatinine	0.87 (0.70, 1.30)	0.82 (0.71, 1.05)	0.3
Iron	78 (41, 106)	94 (40, 144)	0.021
Ferritin	355 (161, 863)	295 (80, 905)	0.3

Below is a demonstration of the use of KNN imputation.

```
hcc_knn <-  
  impute_knn(hcc, . ~ target_cat, pool = 'multivariate', k = 5)
```

So far, several methods of imputation have been demonstrated. It would be interesting to see how the summary statistics are affected after imputation. For this, the KNN imputation is used, after which the bivariate analysis of the risk factors is performed. The results are shown in Table 18.5.

## 6.7 Multiple Imputations

Finally, let us discuss multiple imputations. A missing value is imputed multiple times in multiple imputations [8, 9]. Thus, the process creates multiple independent data sets. Suppose there are five columns and one hundred records in the data. Some

of the records have missing values. In multiple imputations, if imputation is repeated, say five times, then five copies of the original data (with missing values) will be created. Each data set will then be imputed independently of other data. At the end of imputation, there will be five data sets with completed records (due to imputation). These data sets have identical values for the observed entries. They differ only on the imputed data values. The amount of variation in the imputation would indicate uncertainty about the imputed values.

Multiple imputations separate the imputation and analysis steps into two sets of problems. In the first step, multiple imputed data sets are created. In the second step, statistical analysis is performed on each data set. Finally, the results are pooled using Rubin's rule [8] into final point estimates and standard errors. Under the appropriate conditions, the pooled estimates are unbiased with desired statistical properties [1].

As discussed before, a significant advantage of multiple imputations is appropriately addressing the inherent problem with a single imputation. Most single imputation techniques produce biased estimates that underestimate the standard error.

### 6.7.1 Multiple Imputations Demonstration

For demonstration, only a few variables are selected to model one-year survival. Consider the following variables for the demonstration: gender, symptoms, age\_at\_dx, total\_bill, and iron. First, a binary variable from the categorical target is created. This is to confirm the requirements of `glm()` function. Here, a binary logistic regression model is fitted to predict one-year survival using the above predictors. For multiple imputations, `mice` library was used [10].

```
options(digits=4)

hcc_glm <- hcc %>% mutate(
  target_bin = if_else(target_cat == 'Survived', 1, 0)
) %>% select(target_bin, gender, symptoms, age_at_dx,
              total_bill, iron)

fit_obs <- glm(target_bin ~ gender + symptoms + age_at_dx
               + total_bill + iron, data = hcc_glm,
               family = 'binomial')

coef(summary(fit_obs, digits = 3))

##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  2.60422   1.576293  1.6521 0.098511
## gender       0.28408   0.820550  0.3462 0.729186
## symptoms    -0.79613   0.581890 -1.3682 0.171256
## age_at_dx   -0.03992   0.020565 -1.9412 0.052230
```

```

## total_bill -0.38425  0.140028 -2.7441  0.006068
## iron        0.01611  0.006346  2.5382  0.011142

options(digits=4)
library(mice)

## 
## Attaching package: 'mice'

## The following object is masked from 'package:stats':
## 
##     filter

## The following objects are masked from 'package:base':
## 
##     cbind, rbind

hcc_imp <- mice(hcc_glm, seed = 19, m = 20, print = FALSE)
fit_imp <- with(hcc_imp, glm(target_bin ~ gender + symptoms +
                             age_at_dx + total_bill
                             + iron, family = 'binomial'))

summary(pool(fit_imp), digits = 2)

##          term estimate std.error statistic      df p.value
## 1 (Intercept) 2.22391  1.11918    1.987 138.50 0.048886
## 2 gender      -0.15953  0.51468   -0.310 119.84 0.757125
## 3 symptoms    -1.04389  0.44604   -2.340 122.69 0.020881
## 4 age_at_dx   -0.02471  0.01469   -1.682 142.28 0.094802
## 5 total_bill   -0.21527  0.08838   -2.436 70.03 0.017409
## 6 iron         0.01617  0.00594    2.722 42.33 0.009376

```

Recall that there were missing values in some of the predictors in the model above. First, a logistic regression model was fitted. By default, R's `glm()` function deleted the cases listwise and then fitted model. Next, imputed data sets were created using `mice()` function. Here, the `seed = 19` was used for reproducibility, and `m = 20` indicates the number of copies of the imputed data sets to create. Finally, the model was refitted utilizing the imputed data sets using the `with()` function. The results are then pooled.

The pooled results show a slightly lower standard error of the estimates than the complete case analysis.

One common question on multiple imputations is how many data sets to generate. It depends on the data itself, and the researcher may have to do some trial and error. Often a small number of repetitions ( $m$  as small as 2) could be enough. When  $m$  is large, the standard errors of the mean will reduce. Further, if interest is focused on point estimates alone, then " $m$ " should be as high as possible. Depending on the size of the data, it might be computationally intensive.

Finally, multiple imputations solve a problem when the researcher accepts pooled estimates. This is typically the case when some modeling is involved. As for descriptive summary statistics, the software can perform various statistics analyses after multiple imputations. After multiple imputations, the `mice` package in R can do pooled one- and two-sample t-test. For pooling Chi-squared tests, `miceadds` R

library can be used. Mice can also pool results for binary logistic regression (demonstrated above) and Cox proportional hazards model.

## 7 Further Practice

To test your understanding of missing values, their types, and how to handle missing values at the analysis stage, please attempt these questions. Solutions to these questions are provided at the end of the section.

1. Which of the following is the correct definition of missing completely at random (MCAR)?
  - (a) When the probability of missing value on a variable  $Y$  is related to the variable itself
  - (b) When the probability of missing value of a variable  $Y$  is completely unrelated to the missing value in another variable in the data
  - (c) When the probability of missing value on a variable  $Y$  is related to a variable that was not measured
  - (d) When the missingness on  $Y$  is unrelated to the most
2. Which of the following is the correct definition of missing completely at random (MCAR)?
  - (a) When the missingness is unrelated to the variable itself but somewhat related to the other observed data
  - (b) When the probability of being missing is affected by factors within groups of observed data
  - (c) When the missingness is systematically related to the unobserved data. That is, the researcher has not measured some factors or variables that may have caused missing values
  - (d) The probability of missing value on a variable  $Y$  is related to all the observed variables
3. Read the following scenario and identify what type of missingness has occurred:  
A measurement scale has produced missing data only when the scale was placed on an uneven surface
  - (a) Missing at random
  - (b) Missing completely at random
  - (c) Missing not at random
  - (d) None of the above
4. Suppose a researcher has performed multiple imputations with 10 repeats. What is the most appropriate next step to analyze the imputed data sets?
  - (a) The researcher would randomly select one of these data sets and perform analysis on it

- (b) Analyze all of the imputed data sets and average the results to create final results
  - (c) The researcher may or may not analyze all the imputed data sets. But should they do it, one random selected result would be good enough
  - (d) Depending on the objective of the analysis, the multiple imputed data sets would be analyzed separately, and the results would be pooled together to create a single estimate
5. Suppose multiple imputations are performed with a sufficiently large number of repeats (say  $m = 30$ ). In that case, the estimated standard error would likely be lower than the standard error obtained from the observed data only (not imputed, but using complete cases). True/False
6. For multiple imputations, if the number of multiply imputed data sets is too small (say  $m = 2$ ), the result may be affected due to statistical inefficiency. True/False
7. Which of the following is the most appropriate statement about identifying missing values
- (a) Missing values do not pose a threat if missingness is in the outcome variable
  - (b) Missing values only matter if there is a large number of missing values otherwise, we can simply ignore them
  - (c) It is all right to perform listwise deletion if the missingness is about 10%
  - (d) Missing values should be explored at the exploratory data analysis (EDA) phase, and an appropriate method be used to handle them properly
8. Which of the following is an appropriate strategy for imputing missing values in a categorical variable?
- (a) Impute them with a modal value if the missing percentage is negligible
  - (b) Convert the category to numeric values and impute them with the mean of the variable
  - (c) Missing values in a categorical variable would not affect the analysis
  - (d) Let the statistical software handle it during analysis
9. Predictive mean matching is applicable for categorical data only. True/False
10. True or False: When imputing values using predictive mean matching, it is possible to have meaningless imputations (for example, negative height measures). True/False
11. Which is the correct statement about Rubin's rules?
- (a) Rubin's rule specifies how to evaluate the effect of missing values in the data
  - (b) the rules help researchers selectively choose pooling methods to combine the estimates
  - (c) the rules are proposed for pooling the parameter estimates as applied to estimating regression coefficients, and standard errors, for example
  - (d) Rubin's rules are a set of rules to determine how to perform imputation

12. Consider a data set with 50 rows and 10 variables. If three variables have missing values on the same record, how many rows will be dropped if we perform listwise deletion?
- (a) 1
  - (b) 3
  - (c) 10
  - (d) none

### Answer Keys

- 1. b
- 2. c
- 3. a
- 4. d
- 5. True
- 6. True
- 7. d
- 8. a
- 9. False
- 10. False
- 11. c
- 12. a

### References

1. Van Buuren S. Flexible imputation of missing data. Boca Raton, Florida: CRC Press; 2018.
2. Little RJ, Rubin DB. Statistical analysis with missing data. Hoboken, New Jersey : Wiley; 2019.
3. Santos MS, Abreu PH, García-Laencina PJ, Simão A, Carvalho A. A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. *J Biomed Inform.* 2015;58:49–59.
4. Rubin DB. Inference and missing data. *Biometrika.* 1976;63:581–92.
5. Andridge RR, Little RJ. A review of hot deck imputation for survey non-response. *Int Stat Rev.* 2010;78:40–64.
6. van der Loo M. Simputation: simple imputation, 2022.
7. Gower JC. A general coefficient of similarity and some of its properties. *Biometrics.* 1971;27:857–71.
8. Rubin DB. Multiple imputation for nonresponse in surveys. Hoboken, New Jersey: Wiley; 2004.
9. Rubin DB. Multiple imputation after 18+ years. *J Am Stat Assoc.* 1996;91:473–89.
10. Van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. *J Stat Softw.* 2011;45:1–67.

# Chapter 19

## Artificial Intelligence and Machine Learning



Hamidreza Moradi

### Learning Objectives

After completing this chapter, you will be able to:

- Understand and implement machine learning models for regression, classification, and clustering tasks
- Utilize both machine learning and deep learning to devise a predictive model
- Apply concepts of computer vision and natural language processing

## 1 Introduction

Machine learning (ML) has gained a lot of attention in recent years, and many tools and libraries have been developed to help enthusiasts. Devising a predictive model requires a good understanding of the underlying concepts and hands-on experience with at least a library or package for developing models. In the first section, we aim to discuss traditional ML methods and the libraries to be utilized for developing a prediction model. The models we considered were regression, classification, and clustering. In regression, we aim to predict continuous values, while with classification, the goal is to predict data labels assigned to data instances. And clustering tries to find a natural grouping in the data. In the second section, we will start by discussing the basic building blocks of a neural network (NN) model. Then, we develop regression and classification models using NNs to get familiar with how the same task can be achieved using traditional and new approaches. Next, we will review more advanced examples that can be better addressed using NNs, computer vision, and natural language processing (NLP). In computer vision, convolutions will be introduced with examples, and we will see how they can be combined to extract features from an image. In NLP, we will address how we can make words and sentences interpretable and understandable by machines. Finally, the last two sections

---

H. Moradi (✉)

Department of Computer Science, North Carolina Agricultural and Technical State University, Greensboro, North Carolina, United States  
e-mail: [hmoradi@ncat.edu](mailto:hmoradi@ncat.edu)

will briefly discuss how we can make ML models interpretable and recent advancements. All code examples are provided here in Python, as it is the preferred language for both ML and deep learning (DL) for students and engineers in academia and industry. Simplicity, flexibility, platform independence, being open-source, and the existence of powerful libraries/frameworks supported by industry leaders are just a few reasons for its popularity and attention.

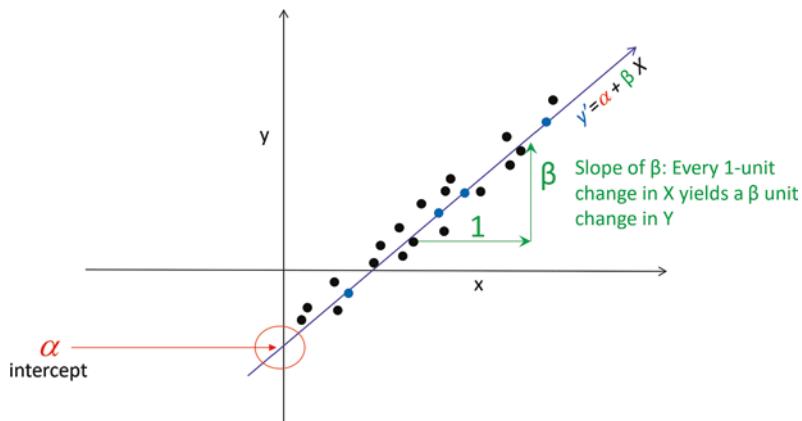
## 2 Machine Learning

In artificial intelligence (AI), computer systems are programmed to mimic human behaviors to devise intelligent machines. This required writing programs with the rules needed to make decisions based on the inputs to the system. However, with increased data and computation power, research studies investigate algorithms that, given the inputs and expected outputs, can automatically infer the rules needed, learning from examples. This marked the beginning of the ML era, a subset of AI. In the following subsections, we will discuss some of the ML algorithms that can be used for inferring the rules and making a prediction for the expected output.

### 2.1 Regression

Regression is one of the early forms of ML models that establishes a linear relationship between a dependent variable and one or more independent variables, usually called features. Linear regression can be graphically presented using a straight line and can be used to predict continuous values. In its simplest form, it can be represented in the form of  $y' = \beta + \alpha X$ , where  $\alpha$  is the weight for the input feature  $X$ ,  $\beta$  is the intercept for the line, and  $y'$  is the predicted value. With ML, the aim is to find  $\alpha$  and  $\beta$  using a dataset of samples. Figure 19.1 shows the data samples used to find a regression line with black dots and the devised regression line in blue. The blue dots on the regression line show a couple of predicted values for the corresponding input ( $X$ ) on the regression line.

The error between the actual observations and their predicted values should be minimized to find the regression line with the best fit. As a result, we are interested in finding the values for  $\alpha$  and  $\beta$  that minimize the prediction error  $\varepsilon = y_i - y'_i$  for all the observations, where  $y$  is the actual observation,  $y'$  is the predicted value, and  $i$  is the observation number in the sample dataset. We need to define and minimize a cost (loss) function here. The cost function for linear regression can be defined as  $\mathcal{L} = \sum (y_i - y'_i)^2$  to penalize large errors and should be minimized by an ML algorithm.<sup>7</sup> In regression, when more than one independent variable exists, the formula can be extended to the form  $y' = \beta + \alpha_1 X_1 + \alpha_2 X_2 + \dots + \alpha_n X_n$ , where  $n$  is equal to the number of features for each observation. Then, the ML algorithm needs to find the best  $\alpha_i$  that minimize the cost function.



**Fig. 19.1** Regression line with actual and predicted observations

Let's see how to develop an ML model. To train a linear regression model in Python [1], Scikit-Learn [2] library provides many preimplemented functions. These functions can be easily utilized to fit a model to a dataset. After installing the latest versions of Python (v3.8) and Scikit-learn (v1.0.2), we need to import the linear models from the library into the working environment in the code editor of choice.<sup>1</sup> Scikit-Learn also provides many small toy datasets [3] to help users try and learn about the implemented functionalities. Here we will import and use the diabetes dataset. Features include patients' baseline measurements such as body mass index, age, blood, and glucose level to predict a quantitative measure of disease progression a year after the baseline. As the outcomes considered for patients are integer numbers (between 25 and 346), we will use linear regression for modeling and prediction.

```
from sklearn import linear_model
from sklearn import datasets
diabetes = datasets.load_diabetes()
input = diabetes.data
outcome = diabetes.target
```

### Train-Test split

In ML, the goal is to train a predictive model and then evaluate the accuracy of an unseen data set. This gives us a measure of how the model may perform in the production environment. To achieve this goal, we can split our data into nonoverlapping train and test subsets of 80% and 20%, respectively.

<sup>1</sup>PyCharm or Jupyter Notebook are recommended.

To have the most accurate model and find the weight for each feature representative of its importance, it is best to standardize the features using a  $z-score$ . This will result in inputs to the model on the same scale. But to normalize the data, we should consider the test set as unseen. As a result, we need to learn the parameters required to standardize the data from the training set and use the same parameters to transform the test data.

```
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
in_train, in_test, out_train, out_test =
train_test_split(input, outcome, test_size=0.20)
scaler = StandardScaler()
scaler.fit(in_train)
in_train = scaler.transform(in_train)
in_test = scaler.transform(in_test)
```

In the above code block, the “scaler.fit” step will learn the required parameters to standardize each feature separately. Then, using the learned parameter from the train set (in train), both the training and test datasets are transformed.

Now, we can train the model using a training set to predict the output of the test data and evaluate its accuracy.

```
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error
reg = LinearRegression().fit(in_train, out_train)
pred_test = reg.predict(in_test)
print(mean_squared_error(out_test, pred_test))
```

Here we used mean squared error to measure accuracy, but many other metrics can be used. The most popular ones that readers are encouraged to review are  $R$  square, mean absolute error, and mean relative absolute error.

Simple linear regression, as used above, does not always provide the best results possible. The outliers greatly impact it and may learn random variations within the training data, causing an issue known as model overfitting. The cost function can be adjusted to address these limitations by adding a penalization term to consider the features’ weight. This will result in three derived regression models with differences in their regularization terms: Lasso, Ridge, and Elastic-Net.

In Lasso regression, weights will be used as the penalization term in the loss function formed as  $\mathcal{L} = \sum (y_i - \hat{y}_i)^2 + \lambda \sum |w_i|$ , where  $\lambda$  is a hyperparameter to tune and  $w_i$  is the weight for the feature  $i$ . Using the sum of the absolute values of weights, known as L1 loss, will result in smaller weights closer to zero. This will cause sparsity in the weights and is used as a feature selection method. Below are the codes needed to import and train Scikit-Learn’s Lasso regression [4] to make predictions for the discussed dataset.

```
from sklearn.linear_model import Lasso
reg2 = Lasso().fit(in_train, out_train)
pred2_test = reg2.predict(in_test)
print(mean_squared_error(out_test, pred2_test))
```

**Note 1** For simplicity, here we trained Lasso regression using the default parameters. However, given that we now have  $\lambda$  as a hyperparameter, we may need to tune it and find the best value for it. One approach would be to split the training dataset into new training and validation sets. Models can be trained on the new training set using different values of  $\lambda$ , and results will be evaluated on the validation set to find the best value. Then, the model with the best  $\lambda$  value can be used to evaluate the accuracy of the test dataset. This approach will allow us to fine-tune the model's hyperparameter and use the best model for the test dataset without data leakage.

Ridge regression will similarly penalize the weights of the features in the loss function. However, the final loss function will be in the form of  $\mathcal{L} = \sum(y_i - \hat{y}_i)^2 + \lambda \sum w_i^2$ . Using the sum of squares of weights in the loss function, known as L2 loss, will penalize the larger values with higher intensity, resulting in a more uniform range of weights. Ridge regression is useful when there are too many features or when features have a high degree of multicollinearity. The code below utilizes Ridge regression for modeling and prediction.

```
from sklearn.linear_model import Ridge
reg3 = Ridge().fit(in_train, out_train)
pred3_test = reg3.predict(in_test)
print(mean_squared_error(out_test, pred3_test))
```

Finally, Elastic-Net combines the two aforementioned regularization terms to form the loss function  $\mathcal{L} = \sum(y_i - \hat{y}_i)^2 + \lambda_1 \sum |w_i|^2 + \lambda_2 \sum w_i^2$ . Here, we can take benefit from both the regularization terms discussed for Lasso and Ridge regression. However, the hyperparameter tuning for both  $\lambda_1$  and  $\lambda_2$  is needed.

```
from sklearn.linear_model import ElasticNet
reg4 = ElasticNet().fit(in_train, out_train)
pred4_test = reg4.predict(in_test)
print(mean_squared_error(out_test, pred4_test))
```

**Note 2** There are different methods of hyperparameter tuning. Grid search will evaluate all possible combinations of hyperparameters with their provided search space to find the best subset. This search method implemented in Scikit-Learn [5] is recommended for a small set of hyperparameters. The search space will grow exponentially with an increase in the number of hyperparameters or their corresponding search space. Random search [6] implementation will use a random combination of values for each hyperparameter for a preset number of iterations and report the best combination. Although random search has gained a lot of interest and has proven to

be efficient, more advanced techniques exist to better utilize any correlation between the accuracy achieved and the parameters tried to make a more informed decision about the next set of values to try. The HyperOpt [7] library is an example that utilizes Bayesian optimization for hyperparameter tuning.

## 2.2 Classification

In classification, the goal is to use ML algorithms and assign a class label to the input examples. For instance, class labels can be patients' death or discharge outcomes at the end of a treatment or negative or positive blood test results. These examples are called binary classification since one of the outcomes can happen at a time and can be coded as a zero or one output for a model. It is possible to have more than two outcome classes as well, where instances can still belong to only one class, usually referred to as multiclass classification. In the case of multilabel classification, input examples can have more than one class label.

There are many ML algorithms designed to address each task mentioned above. Some can only perform simple binary classification, while many are inherently capable of multiclass or multilabel classification. Still, advanced techniques can be incorporated to use a simple binary classifier and form a multiclass or multilabel classifier.

Logistic regression is a simple binary classifier that essentially passes the output of a linear regression model, here  $f(x)$ , through a sigmoid function  $p(x) = 1/(1 + e^{-(x)})$ . This will result in an output with values between zero and one. The final class label can be considered a positive outcome for a predicted value above a specific threshold, while the value below the threshold will be interpreted as negative.

K-Nearest Neighbor (KNN) is a multiclass classifier. It uses a measure of distance<sup>2</sup> (e.g., Euclidean, Hamming, Cosine, Manhattan) to find the K closest neighboring data instances in the training set to a given input sample. Then, the class labels for the determined K-nearest neighbors will be used in a voting scheme to decide the best matching class label. While having a simple algorithm, KNN does not perform well on very large or high-dimensional datasets.<sup>3</sup>

In recent years, tree-based models have shown great efficiency for classification tasks with high prediction accuracy. Tree-based models are based on the simple idea of a decision tree, where a series of conditional steps are taken to make a decision. Figure 19.2 shows an example of a decision tree. However, simplicity comes with disadvantages. Model overfitting arises when the tree fits well to the training data but performs poorly on the testing set. Moreover, considering the order of the conditional steps taken, trees of various shapes will be generated and performed

<sup>2</sup>It is best to use a measure of distance relevant to and representative of available features.

<sup>3</sup>With an increase in data dimensionality, the data points will appear closer together, making this algorithm inefficient. Additionally, pairwise comparison to find the closest neighbors makes the algorithm inefficient in datasets with a large number of instances.

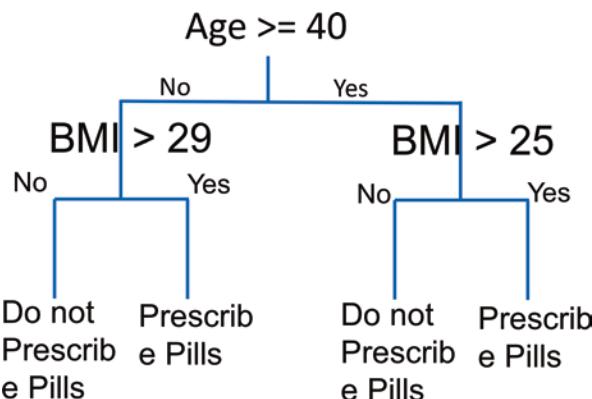
differently. These issues resulted in the invention of two well-known models based on decision trees: Random Forests (RF) and Gradient Boosting Decision Tressses (GBDT).

RF builds a bunch of decision trees independently, each making a simple prediction. Each tree's structure will be randomized and created on top of a bootstrap<sup>4</sup> sample of the training dataset. The final prediction by RF will be the aggregated prediction of all trees. The randomization of the structure and use of bootstrapping have made RF a powerful model resistant to outliers and missing values in the datasets.

### Boosting

To build a more robust model, GBDT models and popular examples (e.g., XGBoost and CatBoost) use an ensemble of weak decision tree predictors. In GBDT, tressses are built iteratively. Meaning each tree is built after the other, and the previous step's output is used in addition to the features as input to the next tree. This will result in each new tree improving on the predictions made by the previous round, improving overall efficiency, a concept called boosting.

Let's get to coding by considering a classification task. For the dataset, we considered using Scikit-Learn breast cancer data [8]. This dataset includes samples of 569 patients with 30 numeric predictive attributes, each labeled as malignant or benign. Like regression, we need to load the dataset, split the data to train and test sets and standardize the features.



**Fig. 19.2** A decision tree to predict patients' need for a prescription

<sup>4</sup>Random sampling with replacement.

```
breast_cancer = datasets.load_breast_cancer()
features = breast_cancer.data
labels = breast_cancer.target
f_train, f_test, l_train, l_test = train_test_split(features,
labels, test_size=0.20)
scaler = StandardScaler()
scaler.fit(f_train)
f_train = scaler.transform(f_train)
f_test = scaler.transform(f_test)
```

For the classification algorithm, we will use the Scikit-Learn implementation of RF [9]. We need to first fit the model to the training data using the “.fit” call. Then the trained model can be used to predict the class labels for the test dataset. Finally, to measure the model’s performance, we can calculate the percentage of correct predictions using the “accuracy\_score” function imported from “sklearn.metrics.”

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import accuracy_score
rf = RandomForestClassifier().fit(f_train, l_train)
pred_rf = rf.predict(f_test)
print(accuracy_score(l_test, pred_rf))
```

We can try different classification algorithms to find the best-performing algorithm for a given dataset. Here, we use the Scikit-Learn implementation of GBDT [10] to evaluate the performance gain achieved using a more advanced implementation of decision trees on the same predictive task.

```
from sklearn.ensemble import GradientBoostingClassifier
gbc = GradientBoostingClassifier().fit(f_train, l_train)
pred_gbc = gbc.predict(f_test)
print(accuracy_score(l_test, pred_gbc))
```

Results show improved accuracy by using the GBDT model over the RF. However, it should be noted that there are hyperparameters for each algorithm that need to be tuned using grid search, random search, or Bayesian optimization techniques, as discussed previously.

**Note 3** For RF, some important hyperparameters to consider are the number of estimators, the maximum depth of trees, and the criterion to measure the quality of splits for each feature. GBDT, while providing a similar hyperparameter to tune, has a few additional unique hyperparameters, such as loss function and learning rate, to consider.

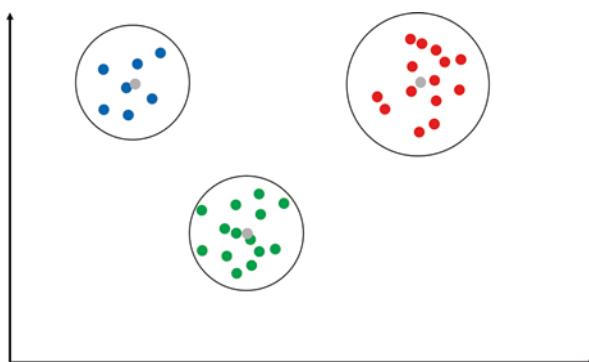
**Note 4** It should be noted that the percentage of correct predictions for a classification task is not the only measure of its performance. The confusion matrix, the area under ROC curve (AUC), and the F1 score are a few others.

## 2.3 Clustering

So far, all the ML algorithms we discussed are considered supervised learning techniques. In supervised learning, the data instances have a class label or a value assigned. As an expected outcome, this value or label will be used to train a model and later need to be predicted for new instances. In contrast, clustering is an unsupervised ML method that involves discovering a natural grouping among the examples. A cluster is an area of densely populated samples or samples closer to each other. Clustering helps us better understand a problem or dataset, group similar data instances, or map new data to an existing cluster in the dataset. Figure 19.3 shows data instances in a two-dimensional space with three clusters. Each cluster is identified with a separate color surrounded by a black circle and a gray dot in the center as a cluster centroid.

K-means and OPTICS are two clustering algorithms we are going to discuss here. K-means requires the number of clusters ( $K$ ) and a measure of distance to be defined and used for calculation. It starts by considering  $K$  random values<sup>5</sup> as cluster centroids. Then, it calculates the distance between each data instance and all the cluster centroids, assigning the closest cluster to it. When the initial clustering happens for all the data instances, K-means will move each cluster centroid to the center of the data instances assigned to it.<sup>6</sup> With the new cluster centroids, K-means will recalculate the distances and assign data instances to the newly formed cluster centers. This process will be repeated until there is no change in the assigned cluster for any data instances or it reaches an identified maximum number of iterations.

While K-means splits the feature space into distinct areas with its measure of distance, OPTICS uses a measure of density and reachability in a provided neighborhood for clustering. Although both clustering algorithms may provide the same results in some cases, the results of applying each clustering algorithm could be



**Fig. 19.3** Three clusters of data in a two-dimensional space

<sup>5</sup>The value considered as  $K$  is a hyperparameter that needs to be tuned.

<sup>6</sup>This will be done by taking the average of each feature for the data instances assigned to the same cluster.

drastically different. Many evaluation metrics are devised for clustering algorithms. However, there is no best or easiest method of comparison. Evaluation of the identified cluster may require controlled experiments or domain expert knowledge. Figure 19.4 shows clusters found by both K-means and OPTICS differentiated by their colors.

To increase the interpretability of our analysis, here we use the Scikit-Learn synthetic dataset generator to create a data set for clustering. The dataset consists of 500 samples in a two-dimensional space, intentionally generated to have five clusters with varied densities, as shown below.

```
from sklearn.datasets import make_blobs
X, y = make_blobs(n_samples=500, centers=5,
cluster_std =1.00)
```

We can now import K-means clustering [11] from Scikit-Learn and apply it to synthetically generate a dataset<sup>7</sup> to form five clusters as we expect. The clusters will be formed using the “.fit” call, and by “.cluster\_centers\_” we can print out the centers calculated.

```
from sklearn.cluster import KMeans
km = KMeans(n_clusters=5)
km.fit(X)
print(km.cluster_centers_)
```

We will use the Seaborn library to generate the figure for the clustered data. Figure 19.5 shows the clusters identified by different colors for the synthetically generated data with five known clusters.

```
import seaborn as sns
sns.scatterplot(x=X[:,0], y=X[:,1], c= km.labels_)
sns.scatterplot(x=km.cluster_centers_[:, 0],
y=km.cluster_centers_[:, 1], c=["black"])
```

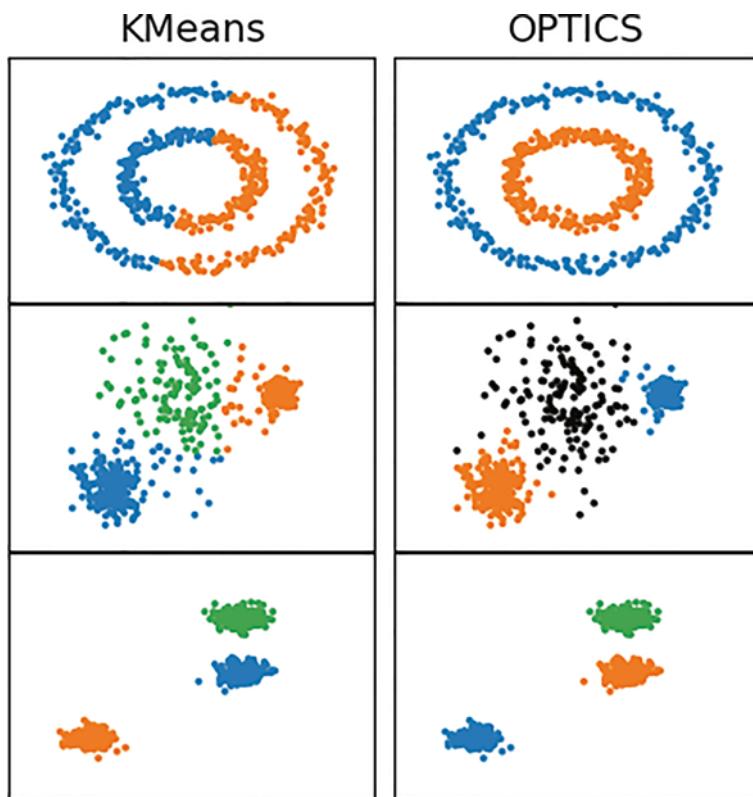
Like K-means clustering, we can apply OPTICS [12] to the same dataset getting the identified clusters and comparing the two algorithms. Figure 19.6 shows the results of utilizing the OPTICS algorithm.

```
from sklearn.cluster import OPTICS
opt = OPTICS()
opt.fit(X)

sns.scatterplot(x=X[:,0], y=X[:,1], c= opt.labels_)
```

---

<sup>7</sup>Forming a train and test splits may not be required as we do not have any assigned labeled to the data.



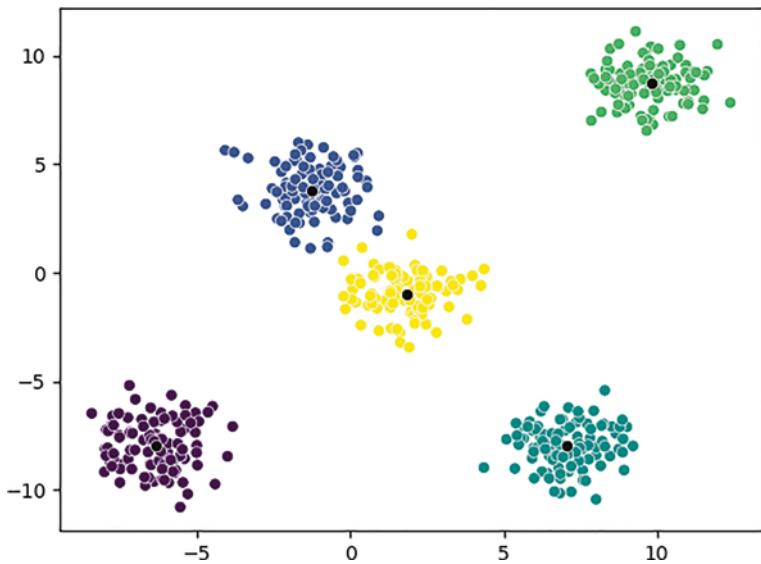
**Fig. 19.4** Clusters found by both K-means and OPTICS

While the results for K-means clustering seem more appealing, if we consider the density-based nature of the OPTICS algorithm, we can conclude that both algorithms are clustering the data perfectly based on their similarity measure. We can observe that K-means separated the two-dimensional space into  $K = 5$  distinct regions with data instances close together. At the same time, optics has found and clustered the data instances with the same density at the center and border of each cluster.

**Note 5** Similar to supervised learning algorithms, unsupervised techniques also have hyperparameters that need to be tuned. For K-means, the number of clusters ( $K$ ) is one of the most important parameters.  $K$  can be found using the elbow method. With the elbow method, clustering will be conducted using different  $K$  values, followed by the calculation of inertia.<sup>8</sup> The  $K$  value as a function of the number of clusters is best where the highest reduction in the inertial is observed, and

---

<sup>8</sup>Inertia measures the sum squared distance between each data point and its assigned centroid.



**Fig. 19.5** K-means clustering and synthetic data

then a plateau is reached.<sup>9</sup> For OPTICS, minimum samples and maximum neighborhood distance are hyperparameters to consider. However, OPTICS presents less sensitivity to hyperparameters than its density-based clustering predecessor, DBSCAN [13].

**Note 6** The silhouette score is another metric for evaluating the clustering algorithms. It measures the degree to which the data points are similar within the assigned cluster compared to neighboring clusters. The silhouette value ranges from  $-1$  to  $1$ , with  $1$  representing the best match and  $-1$  representing the opposite.

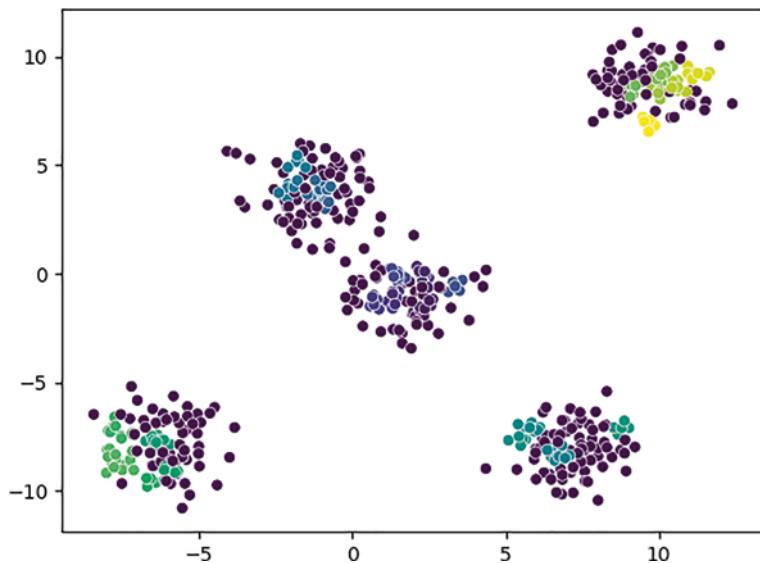
### 3 Neural Networks and Deep Learning

The first interesting and practical demonstration of DL goes back to 1989, when Yann LeCun implemented a NN for handwritten digit recognition. But the capabilities are not limited to computer vision. Now, NNs are used for regression, time-series prediction, object detection and segmentation, robotics, self-driving cars, and even NLP to evaluate text sentiment or generate responses to a question.

NNs, or artificial NNs, are a subset of ML and the heart of DL, inspired by the human brain. It is composed of three or more layers (an input layer, one or more

---

<sup>9</sup>With an increase in the number of clusters, the inertia will constantly decrease, with the lowest inertia achieved with K equal to the number of data instances.



**Fig. 19.6** OPTICS clustering and synthetic data

hidden layers, and an output layer). Each layer is composed of neurons connected to other layers with edges and their associated weights. If the sum of the inputs to a neuron, multiplied by their corresponding weights and added by the neuron's bias value, passes a threshold,<sup>10</sup> the neuron will be activated and pass the signals to the next layer. Generally, NNs start with random weights<sup>11</sup> and zero biases assigned to the edges and neurons. Optimization is needed to generate the required output model to adjust these values. Many efficient optimization algorithms<sup>12</sup> have been devised based on gradient descent and backpropagation, making efficient model adjustments possible. In this order, the inputs will be provided in a forward pass to the model to generate an output. Then the difference between the generated and expected output will be used in a backpropagation step by an optimization algorithm to adjust the weights and biases, reducing the model's error.<sup>13</sup> Figure 19.7 shows a NN with three neurons in the input layer, two hidden layers each with five neurons, and an output layer with two neurons. Here, colored in blue, are the weights connecting input 1 to the next layer of neurons in hidden layer 1.

NNs with enough layers, the correct number of neurons, and the correct settings can virtually simulate any function and provide higher accuracy than traditional models. However, an increase in the model's complexity to achieve higher

<sup>10</sup>This threshold and the degree to which a neuron will be activated are defined by an activation function. The choice of activation function will influence the model's training time and accuracy.

<sup>11</sup>For further details, readers are encouraged to see He and Glorot weight initialization.

<sup>12</sup>Adam optimizer, RMSprop, and AdaGrad are a few to name.

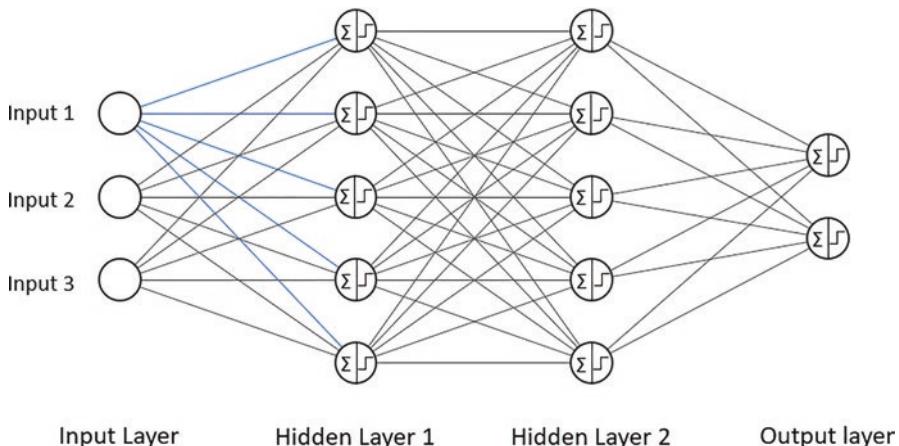
<sup>13</sup>The model's error is calculated using a cost function.

accuracies comes at the cost of requiring more training data. NNs are also better at feature selection than traditional ML models and can accurately learn important features among abundant input. This section will use the Keras v2.10 [14] library backed by TensorFlow v2.8 [15, 16] to develop some of the architectures useful for supervised DL tasks.

### 3.1 NN for Regression

Let's design the architecture of our model to solve the same regression problem we discussed in the first part of this chapter, the diabetes dataset. We have the data loaded, split into training and test sets, and normalized, ready to be used for training a predictive model.

After installing and importing TensorFlow and Keras into the environment, first, we need to specify the type of model implementation we want to use.<sup>14</sup> A sequential implementation is chosen here. Now we can add layers of NN one after the other using the “.add()” command, where the output of one layer will be automatically forwarded to the next layer as input. For the next couple of fully connected layers, a few parameters need to be set. We need to specify the input size for the first layer, the number of neurons, and the activation functions for each layer. We can observe in the code block below that we used the input size of 10, the same as the number of input features, the first and second hidden layers each with 20 neurons,<sup>15</sup> and the



**Fig. 19.7** A simple deep neural network

<sup>14</sup>There are three types of model implementation in TensorFlow: sequential, functional, and subclassing.

<sup>15</sup>The number of hidden layers and the neurons in each layer are hyperparameters that need to be tuned for each model and dataset.

final output layer of size 1 just to predict a single numerical value. The higher the number of neurons in each layer, the more connections with their associated parameters in the model that need to be tuned, increasing the chance of model overfitting. To solve this issue, more training data with a higher computation cost and training time may be required to achieve a good generalizable model.<sup>16</sup>

For each fully connected layer in our model, a few activation functions are available to choose from. Some that we can utilize here are Sigmoid, Tanh, Linear, and Relu. Here we used the Relu function for the first and second hidden layers, as it has proven to be suitable for many predictive tasks and helps train and converge the model faster with fewer epochs.<sup>17</sup> However, the activation function for the output layer is chosen to be a linear function, consistent with the regression task and the range of values to be predicted.

```
from tensorflow import keras

model = keras.models.Sequential()
model.add(keras.layers.Dense(20, input_dim=10,
activation="relu"))
model.add(keras.layers.Dense(20, activation="relu"))
model.add(keras.layers.Dense(1, activation="linear"))
```

The “model.summary()” helps us get a summary of the defined model with the number of trainable parameters within each layer. By compiling the model, we are required to specify a loss (cost) function, an optimizer algorithm, and an accuracy metric. The loss function provides the library with a means to evaluate how much the generated output deviates from the expected output in each training epoch. And the optimizer algorithm tries to minimize the value of the loss function by adjusting the model’s parameters in backpropagation steps. The Adam optimizer is used here due to its accuracy and efficiency. The metrics provided will be used to evaluate the model’s accuracy during training.<sup>18</sup> Finally, the model can be trained for the specified number of epochs by providing the training data and corresponding expected outputs. After completion of the training, the trained model’s accuracy is evaluated by predicting the results for the test dataset.

<sup>16</sup>To prevent overfitting and get a good generalization, the use of the “Dropout” layer is recommended. This technique will randomly deactivate some of the neurons in each training epoch.

<sup>17</sup>Each forward pass to feed the data into the model, getting an output, and its corresponding backpropagation pass to adjust the model are called an epoch.

<sup>18</sup>While in this task, the loss (cost) function and accuracy metric are the same, this is not always the case and usually happens for regression models.

```

model.summary()

model.compile(loss= "mean_squared_error",
optimizer="adam", metrics=["mean_squared_error"])
model.fit(in_train, out_train, epochs=20)

pred_test= model.predict(in_test)
print(np.sqrt(mean_squared_error(out_test,pred_test)))

```

**Note 7** One of the most important hyperparameters in NN is the number of layers and neutrons in each layer. These parameters should be tuned for each predictive task accordingly. Grid Search, Random Search, or Bayes Optimization can be utilized.

**Note 8** While we used well-known metrics of accuracy, optimization algorithm, and cost function, there are many others to try, affecting the final model's accuracy. It is advised to select a few different options and consider experimenting with a number of combinations to achieve the best possible results.

**Note 9** NNs use random weights for model initialization. It is best to make the results replicable by setting a seed value<sup>19</sup> for the random number generator in both the TensorFlow and NumPy libraries.

### 3.2 NN for Classification

We will consider the breast cancer dataset used previously for the classification task. Here, the number of neurons in the last layer should equal the number of classes in the dataset. As a result, each neuron corresponds to one class. During model training and prediction, the neuron in the last layer generating the largest output will be considered the final predicted class label. However, for a binary classification (cancer or not cancer), it is possible to use a single neuron at the output layer. With a single neuron, results can be interpreted as positive if the generated output is above a threshold and negative in reverse. To practice multiclass classification, we will consider two output neurons to simulate a scenario that can be extended to more than two classes.

We use the sequential model implementation here. There are 30 features in the dataset used as inputs to the model (`input_dim`). It is recommended to have more neutrons than inputs in the subsequent layers, with two layers of each 50 neurons.

---

<sup>19</sup>Seed is the value used in computer systems to generate a sequence of pseudo random numbers. By providing an initial input (seed) to a random number generator, the same sequence of random numbers will be generated.

Fully connected layers may overfit, memorizing the random variations in the data and noise instead of learning interactions. Randomly disconnecting some edges between the fully connected layers will improve the model's generalizability. This is achieved using dropout layers after each fully connected layer. In the last layer, we use two neurons equal to the number of classes we are predicting. Moreover, the SoftMax activation function is considered for the last layer, making the summation of all output equal to one<sup>20</sup> and simulating an Argmax Function.<sup>21</sup>

```
model = keras.models.Sequential()
model.add(keras.layers.Dense(50, input_dim=30,
activation="relu"))
model.add(keras.layers.Dropout(0.25))
model.add(keras.layers.Dense(50, activation="relu"))
model.add(keras.layers.Dropout(0.25))
model.add(keras.layers.Dense(2, activation="softmax"))
model.summary()
```

To compile the model for classification, we need to use categorical cross-entropy loss [17]. Moreover, we considered categorical accuracy to compare if the class with the highest predicted probability matches the label provided in the dataset.

```
from sklearn import preprocessing

model.compile(loss= "categorical_crossentropy" ,
optimizer="adam", metrics=["categorical_accuracy"])
```

The model here has two output neurons, and TensorFlow expects the class labels in a OneHot-encoded format for multiclass classification. For each class, we need to have a column, and only one column per data instance can have one value, representing the corresponding class label. The code below will apply the required transformation to both training and test labels.

```
lb = preprocessing.OneHotEncoder(sparse=False)
lb.fit(l_train.reshape([-1,1]))
binerized_l_train = lb.transform(l_train.reshape([-1,1]))
binerized_l_test = lb.transform(l_test.reshape([-1,1]))
```

Now, we need to fit the model with training data.<sup>22</sup>

<sup>20</sup>Converting numbers into a predicted probability distribution.

<sup>21</sup>Argmax function sets the largest predicted probability equal to one and the remaining values equal to zero.

<sup>22</sup>More training epochs are considered here compared to the previous example with more neurons. The number of training epochs is a parameter that needs tuning and should be chosen by considering the model's training loss to prevent overfitting.

```
model.fit(f_train, binerized_l_train, epochs=50)
```

For the final model evaluation, we will predict the test set's class labels and compare them against the actual labels. As a measure of accuracy, we will calculate the percentage of correct predictions using the Scikit-Learn "accuracy\_score" function.<sup>23</sup> One important note is that the model's predictions are probabilities for each class. To convert these to actual class labels, we need to consider the class with the highest probability for each instance as the final prediction. The NumPy Argmax function will look at the predicted probabilities for each instance and return the corresponding column number for which it has the highest probability.

```
from sklearn.metrics import accuracy_score
pred_prob_test = model.predict(f_test)
prediction_test = np.argmax(pred_prob_test, axis=1)
print(accuracy_score(l_test, prediction_test))
```

**Note 10** In this example, all the features are numerical values. Category features should be converted to OneHot-encoded versions before being used as input to the model.

### 3.3 NN for Computer Vision

The field of computer vision is interwoven with Convolutional Neural Networks (CNNs). CNNs have helped computers achieve accuracy above humans in visual tasks. CNNs are used in many tasks, from object detection and classification to self-driving cars and, recently, in many medical imaging domains to detect symptoms and automate the diagnosis process. But what are the CNNs and how a simple CNN can be implemented for a computer vision task are what we will discuss in this section.

In fully connected NNs, each neuron is connected to every neuron in the next layer. This architecture causes a few issues in image processing. If you use fully connected layers for a computer vision task with an image as an input, while the model may learn and predict the assigned classes, shifting or scaling the object in the image would easily affect the final prediction. Since the model has memorized the exact location and size of an object. To address these limitations, we need feature extractors that can learn simple features or patterns in the initial layers of NNs. More complex features and combinations are learned as the data moves along the layers. The first few layers may only learn the horizontal, vertical, and diagonal lines; while moving along the layers, the combination of features in the previous

---

<sup>23</sup>Other metrics to name are the confusion matrix or area under receiver-operating characteristics.

layer can be learned to shape more complex patterns. Generally, a convolution in CNN acts the same as a feature extractor. It will shift over the image and extract features by multiplying with the underlying pixels or values in each step. Figure 19.8 shows an example of a convolutional feature extractor (filter) in CNN. The pixel<sup>24</sup> values for a hypothetical black-and-white<sup>25</sup> image of size 5 by 5 are shown in white with a feature extractor of size 3 by 3 in green. The feature extractor moves over the image calculating the summation of the products. The resulting output will be of size 3 by 3, being forwarded to the next layer.

Convolutional layers are usually followed by a pooling layer, reducing the output size and decreasing the number of parameters required to train in the subsequent layers. A very popular pooling layer is max pooling of size of 2 by 2, convolving similarly over an image, taking the max value of the pixels. Figure 19.9 shows an example of a max pooling filter applied to the final results of Fig. 19.8.

Now that we know the basics of CNNs, let's get to coding and see how we can use layers of CNNs to detect patterns within an image.

Fashion\_mnist is a dataset of images with 10 classes of clothing items. Each image has one layer, as images are black and white. The first step is to load the images from the dataset and reshape them into arrays with four dimensions, as TensorFlow requires. In the reshape function, we can see the requested dimensions in order: the number of training samples (60,000), the height and width of each image ( $28 \times 28$ ), and the number of layers in each image [1]. To normalize the data for image processing, all the values can be divided by the maximum intensity (255), resulting in an intensity value of 0–1. We need to repeat the same process for our test dataset as well.

```
import tensorflow as tf

mnist = tf.keras.datasets.fashion_mnist
(training_images, training_labels), (test_images,
test_labels) = mnist.load_data()

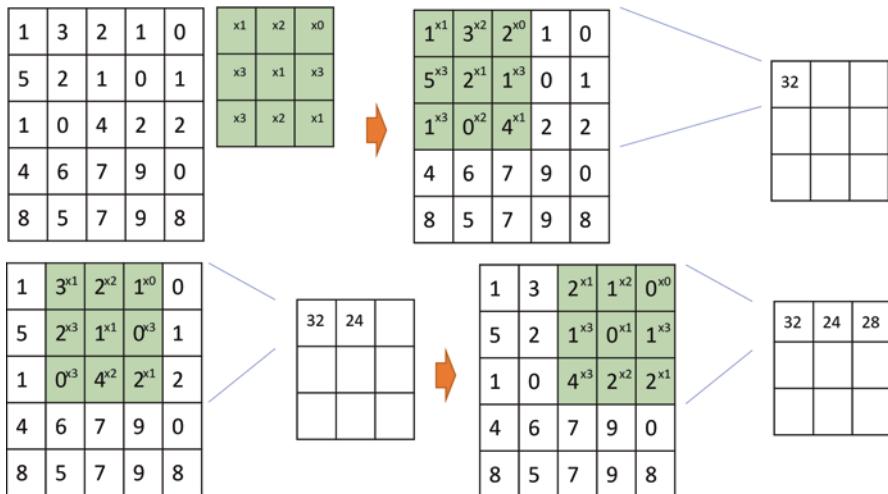
training_images = training_images.reshape(60000, 28,
28, 1)
training_images = training_images / 255.0
test_images = test_images.reshape(10000, 28, 28, 1)
test_images = test_images / 255.0
```

After loading the data, it is time to build our model. Using TensorFlow sequential modeling, we provide the layers in order. In the first layer, we want 64 convolutions with a filter size of 5 by 5 and a ReLu activation function. The input size would be the same as the size of each image, 28 by 28, with 1 layer. The output of this layer

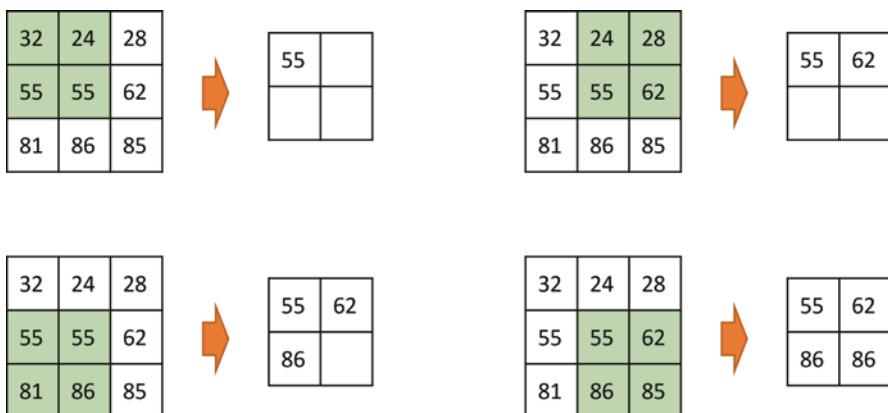
---

<sup>24</sup>The basic unit of a digital image that can be displayed on a digital device or display.

<sup>25</sup>Black and white images can be represented by a single layer, with values presenting the intensity of the light.



**Fig. 19.8** A single convolutional filter convolving over an image



**Fig. 19.9** A max pooling filter convolving over an image

will automatically be forwarded to the next layer. After the convolutional layer, we utilize a pooling layer to reduce the dimensions of the features. We have one more layer of convolutions with 64 filters of size 3 by 3, followed by a pooling layer of size 2 by 2. These layers will extract features from the images. Now, we need to flatten the multidimensional output to one dimension using the “flatten()” function and send it to two fully connected layers for final classification. The final layer needs to have 10 neurons, equal to the number of classes.

```

model = tf.keras.models.Sequential([
    tf.keras.layers.Conv2D(64,(5,5), activation ='relu',
input_shape=(28,28,1)),
    tf.keras.layers.MaxPooling2D(2,2),
    tf.keras.layers.Conv2D(64, (3, 3),
activation='relu'),
    tf.keras.layers.MaxPooling2D(2, 2),
    tf.keras.layers.Flatten(),
    tf.keras.layers.Dense(128, activation=tf.nn.relu),
    tf.keras.layers.Dense(10, activation=tf.nn.softmax)
])

```

To compile the model we just defined, we need to specify the optimizer algorithm, loss function,<sup>26</sup> and accuracy metric. Then, we can get the model summary for parameters and layers and start the training.

```

model.compile(optimizer = 'Adam',
              loss = 'sparse_categorical_crossentropy',
              metrics=['accuracy'])
model.summary()
model.fit(training_images, training_labels, epochs=5)

```

With the trained model, now we can predict the test set and evaluate the accuracy of the model.

```
model.evaluate(test_images, test_labels)
```

### 3.4 NN for NLP

NLP is a branch of ML that allows computers to process and understand text data. This includes, but is not limited to, sentiment analysis, question answering, text summarization, and machine translation. In recent years, DL has revolutionized NLP by developing language models capable of understanding context, trained on millions of documents. In this section, as an introduction to NLP, we will develop a model for sentiment analysis, predicting the positivity and negativity of a text written by a user as a review.

Like many other NLP tasks, the first step is data cleaning and selecting a method for sentence representation. Data cleaning (depending on the task) refers to removing punctuation, extra spaces, HTML tags in the text, emojis, hyperlinks, and stop words. There are quite a few approaches to representing the sentences. One of the

---

<sup>26</sup>The use of sparse categorial cross-entropy loss would eliminate the need to provide the one hot encoded version of class labels.

old methods is a one-hot representation, with one column per word in the dataset. If a word appeared in the sentence, its corresponding column would have a value of 1 and otherwise be zero. While this helps to capture sentiments and represent words in a sentence, the words' order and appearance together will be removed, resulting in information being lost. Using bi-gram and tri-gram can help capture words that appear together and in sequential order. However, this method also suffers from not utilizing the context in which the words have appeared.

For better representations, instead of a sparse matrix with columns representing the words, using an n-dimensional vector to represent each word is a better approach. This n-dimensional representation should encode the word meaning, referred to as embedding. With this approach, we can replace words with embeddings and represent sentences with word-length sequences. There are many pre-trained embeddings<sup>27</sup> to utilize. However, embeddings can be directly learned from a dataset as well.

The dataset we will use here has words pre-converted to unique integer values for simplicity. Integer representations are not recommended as they are randomly chosen. Using an embedding layer in TensorFlow, we can learn the best representation from the dataset for the integer representation provided here.

First, we load the dataset from TensorFlow. “Num\_words” limits the TensorFlow to represent the top 10,000 words based on the frequency of their repetition, considering others as unknown (oov\_char).<sup>28</sup> “Maxlen” instructs the library to truncate any sentence longer than 512 to the same size. Shorter sentences will be padded by zero, making all the representations the same length.

```
import tensorflow as tf

(x_train, y_train), (x_test, y_test) =
tf.keras.datasets.imdb.load_data(num_words=10000)

x_train =
tf.keras.preprocessing.sequence.pad_sequences(x_train,
maxlen=512)
x_test =
tf.keras.preprocessing.sequence.pad_sequences(x_test,
maxlen=512)
```

Now we need to define our network. The “input” variable is a placeholder representative of the sentence that will be passed to the model. The first layer would be the Embedding layer. TensorFlow has recently added this capability, where the model can learn the best representation of the words based on the input dataset. This helps to use representations that accurately encode the word's meaning instead of a randomly generated integer by the data loader. The first value (10000) provided to

---

<sup>27</sup>Glove and word2vec are two of the statically generated word embeddings.

<sup>28</sup>This helps to remove the words that have a very low frequency of repetition.

the embedding layer will represent the expected number of vocabulary words, and the second is the number of output dimensions for representing each word.

```
inputs = tf.keras.Input(shape=(None,), dtype="int32")
x = tf.keras.layers.Embedding(10000, 100)(inputs)
```

We need NN models capable of understanding the sequential nature of words in a language. Although using fully connected layers may be possible, they could not understand the context in which the words appeared accurately. Using recurrent NNs (RNNs), not only are the neurons in each layer connected to the next layer, but there is also a hidden state connecting the neurons within the same layer. These connections are usually unidirectional.<sup>29</sup> But the unidirectional connections can utilize the context information of the previous words for representation. Here, we use a bidirectional implementation of the RNN called Long Short-Term Memory (LSTM) networks to address this limitation. The bidirectional implementation helps information flow from the beginning to the end of a sentence and in reverse, while the LSTM model helps better understand long sequences of words. Finally, the last layer has a neuron as a final classifier, predicting the sentiment for this binary classification task.

```
x = tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(64,
return_sequences=True))(x)
x =
tf.keras.layers.Bidirectional(tf.keras.layers.LSTM(64))(x)
outputs = tf.keras.layers.Dense(1,
activation="sigmoid")(x)
```

The final step would be setting the input and output of the model, followed by compilation and evaluating the model's performance.

```
model.compile("adam", "binary_crossentropy",
metrics=["accuracy"])
model.fit(x_train, y_train, batch_size=64, epochs=3) # ,
validation_data=(x_test, y_test)

model.evaluate(x_test, y_test)
```

---

<sup>29</sup>From the first neural to the last, or from the beginning of the sentence to the end of the sentence.

## 4 Recent Advancements

While here, we provided a couple of source codes and discussed the implementation details. Examples are countless. Applications of ML and DL are not limited to regression and classification tasks or sentiment analysis. In computer vision, with the help of DL, machines can find an object of interest within an image (object detection) and provide an accurate bounding box for it (object localization). They are helping automate the disease diagnosis processes and assisting doctors by proposing areas that might need more attention. Using image segmentation techniques in computer vision, we can provide an exact area where an abnormality is observed with pixel-level accuracy (object segmentation). Studies now focus on utilizing DL to increase the quality of MRI and X-ray images. In NLP, we used the patients' historical notes in EHR to provide clinicians with a summary or even propose a discharge summary. They are helping to improve clinical documentation and saving clinicians' time.

## 5 Models' Interpretability

A black-box ML model usually focuses on predicting outcomes, but little insight is available beyond the predictions. However, recent years have witnessed numerous advances in producing robust and interpretable insights from complex machine-learning models. Some of the examples are the Grad-CAM [18, 19], LIME [20, 21], and SHAP [22, 23] libraries utilized in many applications. The most popular library, SHapley Additive exPlanation (SHAP), is based on the game theoretic approach, which has gained a lot of attention in many domains. SHAP provides insightful interpretations of a complex ML model with high accuracy and robustness, close to human interpretations. The generated SHAP values for input features of an ML model can be used to assess the effect of the inputs on the final model's prediction. There are abundant examples of applications in many domains, including tabular data modeling, text classification, question answering, image processing, and genomics.

## 6 Further Practice

1. What is the cost function?
2. Why do we need to split the data into train and test datasets?
3. Why do we need to standardize the input data?
4. What are the metrics to evaluate the accuracy of the linear regression model?
5. What are the differences between Lasso and Ridge linear regressions?
6. Why do we need to tune the hyperparameters on a validation set?

7. Which terms are different in the L1 and L2 losses?
8. What are binary, multiclass, and multilabel classifications?
9. What are RF and GBDT models, and how do they differ from Decision Trees?
10. What are the differences between K-means and OPTICS clustering algorithms?
11. What should be the input and output sizes of a NN for a given predictive task?
12. How the model generalization can be improved in fully connected NNs?
13. What should be the order of dimensions as input for image processing in TF?
14. Why do we need to use a pooling layer in CNN?
15. What are pre-trained embeddings, and how they can be utilized for an NLP task?

### Answer Keys

1. A function that determines how well an ML model is performing
2. Because the model will be evaluated on the same data it has seen during training. As a result, an evaluation would not be representative of real-world performance with unseen data.
3. To have the same scale for all the data. Many ML algorithms are sensitive to the data scale and may find unrealized coefficients.
4. Mean squared error, mean absolute error, and  $R$ -squared.
5. Lasso cost function will result in some coefficient closer to zero and act as a feature selector, while Ridge will result in a coefficient more uniform and works better with multicollinearity.
6. Before applying models to the actual test dataset and evaluating the performance, hyperparameters should be tuned on an unseen part of the dataset. But this could not be the test set. This is why the training dataset is usually utilized to drive another subset of data to test the model with different hyperparameters to find the best, called the validation set.
7. L1 loss uses the sum of the absolute value of the weight for penalization, while L2 loss uses the sum of squared weights.
8. Binary classification refers to a task that can be seen as a binary outcome positive and negative. Multiclass classification has multiple outcomes in which only one can be true for each instance, while in multilabel classification multiple class labels can be assigned to a single instance.
9. RF uses an ensemble of Decision Trees (DT) using a bagging method by data resampling and combines the result using a voting method. GBDT uses sample weighting or output of previous models to build the next level of predictors, improving modeling accuracy interactively.
10. K-means used a method of distance to find K closest neighbor and uses voting for the final decision, while OPTICS uses density as a measure of clustering using minimum points in a neighborhood in a core distance.
11. Input should be equal to the number of features, size of an image, or length of a sentence based on a task. While output size is equal to the number of classes to predict.
12. Using the dropout layer model generalization can be improved.
13. First is the number of samples, then the size of each sample, and the last dimension represents the number of color channels.

14. The pooling layer reduces the dimensionality of feature, reducing the required computations
15. Pre-trained embeddings provide the representation for each word with meaning encoded in it. Each word should be replaced by a corresponding embedding representation before being forwarded to the model.

## References

1. <https://www.python.org/>
2. <https://scikit-learn.org/stable/>
3. [https://scikit-learn.org/stable/datasets/toy\\_dataset.html.org](https://scikit-learn.org/stable/datasets/toy_dataset.html.org)
4. [https://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.Lasso.html.org](https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html.org)
5. [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.GridSearchCV.html.org](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.org)
6. [https://scikit-learn.org/stable/modules/generated/sklearn.model\\_selection.RandomizedSearchCV.html.org](https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html.org)
7. <http://hyperopt.github.io/hyperopt/>
8. [https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load\\_breast\\_cancer.html.org](https://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_breast_cancer.html.org)
9. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.org>
10. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html.org>
11. <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html.org>
12. <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.OPTICS.html.org>
13. <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html.org>
14. [https://keras.io/getting\\_started/faq/](https://keras.io/getting_started/faq/)
15. <https://www.tensorflow.org/>
16. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16). 2016. p. 265–83. Available from: <https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf>
17. [https://www.tensorflow.org/api\\_docs/python/tf/keras/losses/CategoricalCrossentropy](https://www.tensorflow.org/api_docs/python/tf/keras/losses/CategoricalCrossentropy)
18. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128(2):336–59.
19. <http://gradcam.cloudcv.org/>
20. <https://homes.cs.washington.edu/~marcotcr/blog/lime/>
21. Ribeiro MT, Singh S, Guestrin C. “Why should I trust you?”: explaining the predictions of any classifier. arXiv; 2016. Available from: <http://arxiv.org/abs/1602.04938>
22. <https://shap.readthedocs.io/en/latest/index.html>
23. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, et al., editors. Advances in neural information processing systems. Curran Associates, Inc.; 2017.

# Chapter 20

## A Step-by-Step Guide to Data Analysis Using SPSS: Iron Study Data



Amal K. Mitra

### Learning Objectives

After completing this chapter, you will be able to:

- Get ready for data analysis using SPSS
- Choose statistical tests based on the nature of the variables
- Assess the distribution and normality of data
- Apply appropriate statistical tests that address research questions
- Interpret data outputs

## 1 Introduction

An important first step in data analysis is to get familiar with the study objectives and the data. SPSS (Statistical Package for the Social Sciences) is a powerful data analysis tool. It would be best to familiarize yourself with the arrangement of the variables and commands (or syntaxes) used for statistical tests and graphs. You should also understand the importance of assessing the distribution and normality of the data, without which statistical tests are meaningless and could sometimes be inappropriate. This chapter will help you understand the data analysis process and provide a step-by-step guide to the statistical methods of SPSS and data interpretation. Most importantly, in this chapter, you will know what statistical tests are done, when, why, and how.

If this is your first time using SPSS, you should not worry because it is one of the most user-friendly statistical software programs. Some of the commands depend on which software version you are using. I will use SPSS for Windows, version 28.0, for this lesson. Getting used to one of the latest software versions would be best. If you start SPSS on your computer, you can find two views at the bottom of the screen: the data view and the variable view. Data are displayed like a spreadsheet if

---

A. K. Mitra (✉)

Department of Epidemiology and Biostatistics, Jackson State University, Jackson, MS, USA  
e-mail: [amal.k.mitra@jsu.edu](mailto:amal.k.mitra@jsu.edu)

you look at the data view, and the list of variables is displayed in rows if you use the variable view. In the data view, you may find all the variables in columns at the top and the data arranged under each column in rows.

### **Creating a Data File**

When you create a new data file, there are certain restrictions in naming a variable. For example, (1) you cannot use a number at the beginning of the variable name, but you can use numbers within a variable name; (2) you cannot use hyphenation in between letters or numbers for the variable name, but an underscore sign is allowed; and (3) there is no limit to the characters used for the variable, but the shorter the better. When you use a short variable name, you must label it to recognize the variable at the time of analysis and when reading the results.

### **Naming Variables**

It is often helpful to make the first variable name “ID” or “case number” (in short, “case”). This will help you with using some options, such as “sort cases,” “sort variables,” “merge files,” or “split files.”

#### **Box 20.1**

I advise you to define the variables in the data file only if necessary and only for the purpose of understanding the variable. For example, you are not required to determine the scale of the variable (such as nominal, ordinal, interval, and ratio) or the variable type (such as numeric, string, etc.). The default for the variable type is “Numeric” – you don’t need to change the variable “ID” to numeric, as long as you know that you will not calculate a mean for “ID” or the “case number”. Some people define missing values as “999” or “888”. This is just extra time you spend, which is not always required. You can keep them blank as long as you know the information is missing. When you need to use a code for the missing value, you can come back and define it using an option called “recode.” You can also add a new variable at any time. The following advice is to use multiple people for data entry if it is a large dataset. Either you share a part of the variables (e.g., 60 variables for one person and 50 variables for another person to enter data) or a part of the total subjects (e.g., 200 for one data entry person and another 250 for the next person). You can merge the data into one file for analysis at the end of the data entry. However, you must not forget to use a common variable (such as ID) for each data file when multiple people share data entries. If you think that your control group and the treatment groups should have two separate data files, make sure that the variables in the two files are not different. In our Iron Study data, we had an extra variable called “group.” We coded it as “1 = Control, 2 = Treatment I, and 3 = Treatment II”. For a multicenter study, you can use the same option. All of the variables will be the same, except you will use one extra variable, naming it as “site (or something this nature),” using a different code for each site, and finally merging them. The “merge” option won’t work to combine files if the variables are not identical.

## Using Decimals

As common sense applies, use decimal points as required. For example, if the age in your data has two decimal points, use two decimals; if it is a whole number, there is no need to use decimals. You don't use decimals for ID or case number, which is meaningless.

## Using Important Function Keys

There are many important function key links that you should get familiar with. “File” – you can save or open a data file, syntax, or output. You can use “Utilities” when you want to know about a variable – whether it is a group variable or not; and if it is a grouped variable, what value labels or codes (e.g., 1 for male and 2 for female, etc.) have been used for the variable.

## 2 About the Iron Study

Throughout this chapter, I will demonstrate data analysis using an “Iron Study Data” data file. This study was funded by the Centers for Disease Control and Prevention (CDC). Details of the project objectives, methods and procedures, and results can be obtained from a published source [1].

As mentioned at the beginning, your first step should be to familiarize yourself with the study objective(s), research questions, and variables. The main objective of the Iron Study was to evaluate three treatment options for reducing anemia and improving hemoglobin status in low-income postpartum women through a randomized, controlled, community-based intervention trial. The study was conducted among 959 postpartum women in 11 health clinics in Mississippi. The clinics were randomized to receive one of the three treatment groups:

1. **Control Group** (4 clinics having 249 women): This group was offered selective screening. All women enrolled in these clinics were first screened using the existing method of iron treatment for postpartum women. They are screened if they are at “high risk” based on the following criteria: (1) anemia continued through the third trimester; (2) excessive blood loss during delivery; (3) multiple births; and (4) a previous diagnosis of iron deficiency anemia. Only women at “high risk” were tested for anemia and treated based on the level of anemia.
2. Treatment Group I (4 clinics having 365 women): These women were offered universal screening, meaning all women in these clinics were tested for anemia and treated accordingly. They did not need to wait for a “high-risk” assessment.
3. Treatment Group II (3 clinics having 349 women): All women in this group were under universal treatment, which means women in these clinics were treated with a low dose of iron daily for 2 months, irrespective of their hemoglobin status.

Data on women after delivery were collected at baseline and at a follow-up visit 6 months after delivery. In this study, the women were considered low-income (one of the eligibility criteria) if they were certified for WIC, a special supplemental nutrition program for women, infants, and children.

## 2.1 Major Research Questions

Although the study had several research questions, we will analyze the data to address two major research questions to demonstrate the data analysis that is used in this chapter.

1. Did the hemoglobin levels improve after treatment in the low-income postpartum women?
2. Among the three treatment groups, which group of women improved the most after treatment?
  - (a) Control versus Treatment Group I
  - (b) Control versus Treatment Group II
  - (c) Treatment Group I versus Treatment Group II

## 2.2 Understanding Variables

The Iron Study data comprises 185 variables. Some of the variables that produce quantitative and ratio scale data include age, weight, height, body mass index, hemoglobin status, household members, packs of cigarettes smoked, number of pregnancies, etc. Other important variables, such as treatment groups, race, education, income, and breastfeeding status, are categorical. Knowledge about iron therapy and its benefits is also grouped using a Likert scale. As a data analyst, you should know the abbreviated variable names. For example, the variable name for hemoglobin status at baseline is “hgb,” and the variable name for hemoglobin at follow-up is “hbf.”

### Box 20.2

This information about variables is essential because it guides you in deciding what statistical tests can be applied, when, and for what purpose. For example, hemoglobin status is this study’s most important variable of interest. As a quantitative and continuous variable, serum levels of hemoglobin will generate a mean and standard deviation. To compare hemoglobin status between African Americans and Whites, you should use an *independent sample t-test*. Whereas hemoglobin was measured twice, once at baseline and once at 6 months. To compare the improvement in hemoglobin status among the entire sample, you should use a *paired samples t-test* (also called a *dependent samples t-test*). It will be an ANOVA test if you want to compare the mean hemoglobin status among three or more groups, such as Controls, Treatment I, and Treatment II.

Further details about the statistical tests, steps for using SPSS, and the data output are presented here in this chapter.

## 2.3 *Distribution of Data*

Before initiating inferential statistical tests (such as the *t*-test, chi-square, ANOVA, etc.), you must know the data distribution. If the distribution is normal, you will use parametric tests; if the distribution is nonnormal, you will use nonparametric test statistics. For example, the *t*-test (Student's *t*-test) is a parametric test; the equivalent nonparametric test is called Mann–Whitney *U* test. A paired *t*-test is a parametric test used when the data are normally distributed. If the data are not normally distributed, you should use an equivalent nonparametric paired sample test, the Wilcoxon Signed Rank test, or the Wilcoxon test. There are several methods to assess whether data are normally distributed or not. The methods for assessing normality fall into three broad categories:

- **Using descriptive statistics:** Mean, median, mode, standard deviation, standard error, range, quartiles, percentiles, skewness, and kurtosis. This should be your first step in data analysis. A statistical test list is available under “Statistics” when you use the “Analyze” function key in SPSS.
- **Using graphical analysis:** Visual methods of checking normality are easy to do, but they are often unreliable and do not warrant that the data are normally distributed. The graphical methods include frequency distribution using histogram, Box-Plot display, Stem-and-Leaf display, Q–Q (quantile–quantile) probability plot, etc.
- **Computing normality tests:** Various statistical methods used for data analysis make assumptions about normality. The central limit theorem states that when the sample size is large, violation of normality may not be a significant issue because, with larger samples, the sampling distribution tends to be normal, regardless of the shape of the data [2, 3]. For nonskewed data, when the sample size is as small as 30, and for moderately skewed data, when the sample size is larger than 100, it can approach a normal distribution. Sometimes a transformation such as a logarithm can remove the skewness and allow you to use powerful tests based on the normality assumption [4]. Of the various statistical tests for normality, the most commonly used tests are the Shapiro–Wilk test (S–W test) and the Kolmogorov–Smirnov test (K–S test) [3].

### 2.3.1 Comparing Mean, Median, Mode, and Standard Deviation

Measures of central tendency (mean, median, and mode) give you an idea of where a dataset's “center” value is located. The standard deviation tells you how spread the values are around the mean in the dataset.

Rule 1: The data are normally distributed if mean = median = mode.

Rule 2: If mean > median, the data are skewed to the right or positively skewed.

Rule 3: If mean < median, the data are skewed to the left or negatively skewed.

A high standard deviation shows that the data is widely spread (less reliable), and a low standard deviation shows that the data are clustered closely around the mean (more reliable).

### **Analysis Plan 1: Descriptive Statistics**

#### **Steps for SPSS**

- Go to Analyze
- Choose Descriptive Statistics
- Select Frequencies
- Enter variable(s) in the box: hgb (hemoglobin at baseline)
- Uncheck “Display frequency tables” (if you do not need the frequency distribution)
- Press Statistics: Here, you can select only a few (such as mean, median, mode, std. deviation) or as many statistics as you need to analyze. After you have checked the required statistics, press Continue
- Press OK

### **SPSS Data Output**

<b>Statistics</b>		
<b>Hemoglobin baseline</b>		
<i>N</i>	Valid	559
	Missing	10
Mean		11.734
Median		11.800
Mode		11.8
Std. Deviation		1.7250

### **Interpretation**

In the above example, the data follows Rule 1, where the mean (11.7), median (11.8), and mode (11.8) of baseline hemoglobin (variable name “hgb”) are considered equal, even though the mean is slightly smaller than the median and the mode. This means that hemoglobin status at baseline is **normally distributed**.

In the data output, the standard deviation is 1.725. No universal number determines whether a standard deviation is “high” or “low,” because it depends on the situation. One way to determine if a standard deviation is high is to compare it to the mean of the dataset. A coefficient of variation (CV) is a way to measure how spread out values are in a dataset relative to the mean. The higher the CV, the higher the standard deviation *relative* to the mean. In general, a CV value greater than 1 is often considered high. The CV is calculated as follows [5]:

$$CV = \frac{S}{\bar{x}}$$

where  $s$  = standard deviation and  $\bar{x}$ (x-bar) = sample mean.

$CV = 1.725/11.734 = 0.15$ ; this indicates that the standard deviation is low in terms of the mean.

### 2.3.2 Using Histogram (with a Normal Curve), Skewness, and Kurtosis [6]

A histogram with a normal curve on it will show if the curve takes a Gaussian (or bell-shaped) distribution and if there is any skewness.

#### Skewness (Fig. 20.1)

##### Kurtosis

- A standard normal distribution has a kurtosis of 3 and is recognized as mesokurtic.
- An increased kurtosis ( $>3$ ) can be visualized with a high peak (leptokurtic).
- A decreased kurtosis ( $<3$ ) corresponds to a broadening of the peak (platykurtic).
- Kurtosis = 0, which means platykurtic.
- Kurtosis between  $-2$  and  $+2$  is acceptable (Fig. 20.2).

**Analysis Plan 2** Display a histogram with the normal curve on it. Also, examine skewness and kurtosis.

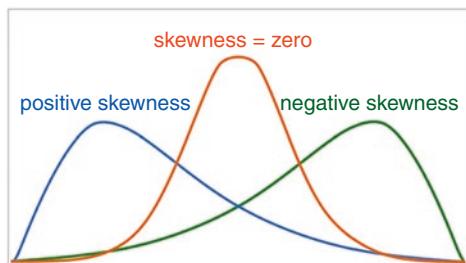
#### Steps for SPSS

- Go to Analyze
- Choose Descriptive Statistics
- Select Frequencies
- Enter variable(s) in the box: hgb (hemoglobin at baseline)
- Uncheck “Display frequency tables” (if you do not need the frequency distribution)
- Select Statistics: check the following boxes – Skewness, Kurtosis
- Continue
- Select Charts: check Histograms, and check the box – Show normal curve on the histogram
- Continue
- OK

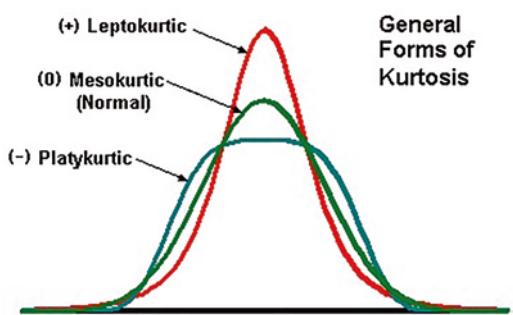
#### SPSS Data Outputs (Fig. 20.3) [1]

Statistics		
<b>Hemoglobin baseline</b>		
<i>N</i>	Valid	559
	Missing	10
Skewness		-0.039

**Fig. 20.1** Types of skewness [6]



**Fig. 20.2** Shapes of mesokurtic, leptokurtic, and platykurtic curves [6]



Statistics	
Std. Error of Skewness	0.103
Kurtosis	-0.211
Std. Error of Kurtosis	0.206

### 2.3.3 Boxplot, Stem-and-Leaf, and Q–Q (Quantile–Quantile) Plot

These methods are used for virtually checking normality distribution.

The boxplot provides a box with the length of the box representing the interquartile range (range between the 25th and 75th percentiles) and a horizontal line inside the box representing the median. The error bars, or whiskers (lines extending from the top and bottom of the box), represent the minimum and maximum values. When the range of the values is greater than 1.5 times the interquartile range, they are considered outliers, and those greater than 3 times the interquartile range are extreme outliers [2].

**Analysis Plan 3** Boxplot display of hemoglobin at baseline (variable name “hgb”) by treatment group (variable name ‘group’, meaning Control, Treatment I, and Treatment II).

#### Steps for SPSS: Boxplot

- Go to Graphs
- Select Legacy Dialogs

- Select Boxplot
- Select Single – Summaries of groups of cases
- Define
- Variable – select hgb (hemoglobin at baseline)
- Category Axis – select group (three treatment groups – control, Treatment I, and Treatment 2)
- OK

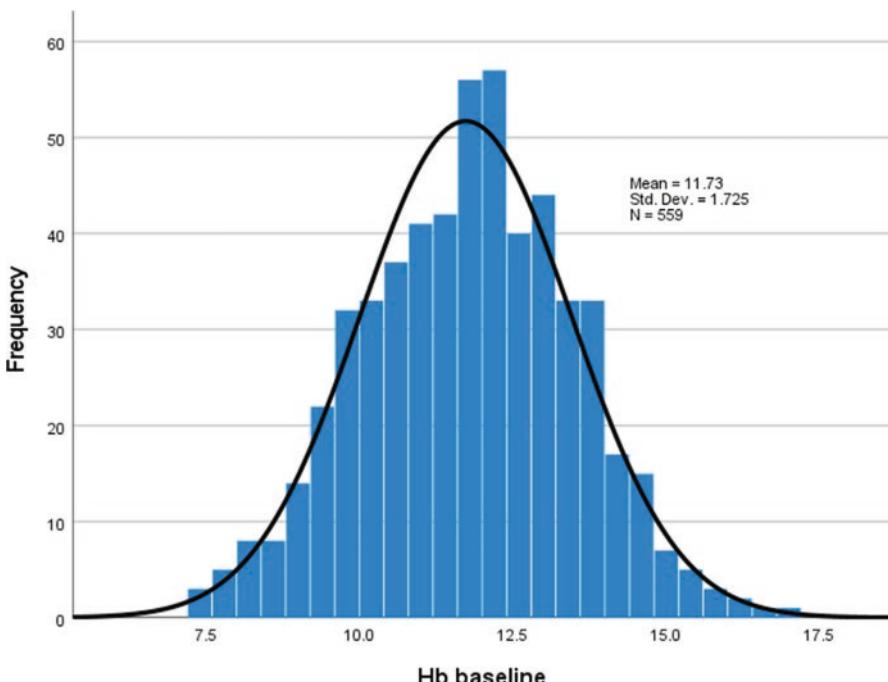
### SPSS Data Outputs (Fig. 20.4) [1]

#### Interpretations

The boxplot distributions for the three groups – Control, Treatment I, and Treatment II – are symmetric, with the median line at approximately the center of the box. The size of the whiskers at the top and bottom of the boxes is almost symmetrical from an arbitrary center of the box, suggesting that the data may have come from a normal distribution.

#### Stem-and-Leaf Display

A stem-and-leaf plot and a histogram are similar. One advantage of a stem-and-leaf plot over a histogram is that the actual data are retained and can be extracted from a stem-and-leaf plot, whereas they can't be extracted from a histogram.



**Fig. 20.3** A histogram showing normal distribution of hemoglobin at baseline. Note: There is no skewness. Kurtosis is low [1]

**Analysis Plan 4** Stem-and-leaf display of hemoglobin at baseline (variable name 'hgb')

#### Steps for SPSS: Stem-and-Leaf Plot

- Go to Analyze
- Choose Descriptive Statistics
- Select Explore
- In the box of Dependent List, enter the variable: hgb (hemoglobin at baseline)
- For Display, select Plots (if you want only plot)
- Check the Plot – the default is Stem-and-leaf
- Continue
- OK

#### SPSS Data Outputs (Fig. 20.5) [1]

##### Interpretations

By looking at the stem-and-leaf plot, you can find the exact hemoglobin levels. For example, the first few hemoglobin levels are 7.3, 7.3, 7.3, and 7.4. The maximum number ( $n = 69$ ) is for hemoglobin levels ranging from 12.0 to 12.4. The display of data looks symmetrical.

#### Q–Q Plot

This is a graphical method for comparing two probability distributions by plotting their quantiles against each other. A Q–Q plot uses the quantiles (values that split a data set into equal portions) of the data set. If the residuals of the data fall along a roughly straight line at a 45-degree angle, then the residuals are roughly normally distributed (Fig. 20.6).

#### Analysis Plan 5 Q–Q plot

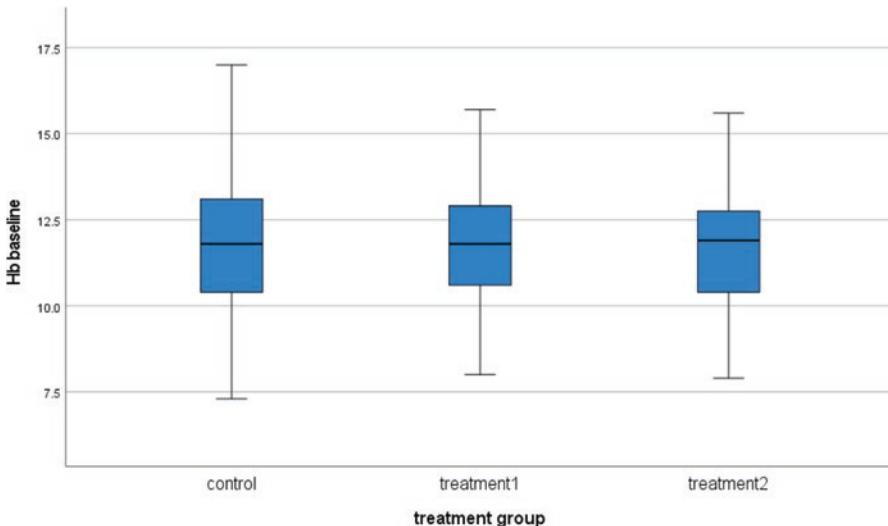
#### Steps for SPSS: Q–Q Plot

- Go to Analyze
- Choose Descriptive Statistics
- Select Q–Q Plots (at the bottom of the list).

#### SPSS outputs (Fig. 20.7)

##### Interpretations

The Q–Q plot shows that our data lie mostly over the straight line, except for a few deviating from the line slightly, especially on the lower tail end. The results indicate that the data are normally distributed.



**Fig. 20.4** Boxplot distribution of hemoglobin at baseline at three treatment groups

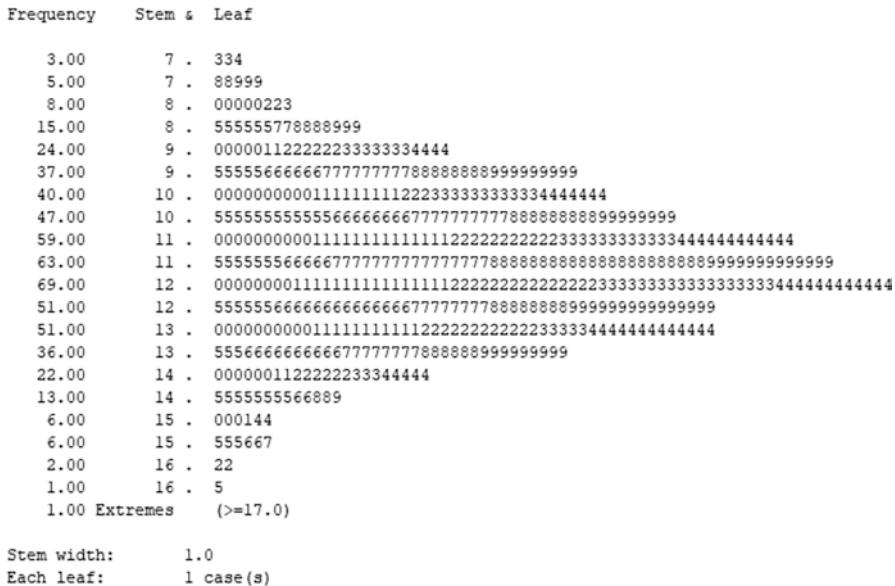
## 2.4 *Shapiro–Wilk Test and Kolmogorov–Smirnov Test of Normality*

There are various methods for normality testing. However, the Shapiro–Wilk (S–W) test is the most popular and widely used method. For small samples ( $n < 50$ ), S–W should be used as it has more power to detect the normality, although it can also handle a larger sample size. Another commonly used test is called Kolmogorov–Smirnov test, which is preferred when the sample size is 50 or more.

**Analysis Plan 6** Shapiro-Wilk test and Kolmogorov-Smirnov test

### Steps for SPSS

- Go to Analyze
- Choose Descriptive Statistics
- Select Explore
- In the box of Dependent List, enter the variable: hgb (hemoglobin at baseline)
- For Display, select Plots
- Check the Plot – the default is Stem-and-leaf (you may uncheck it)
- Check the box – Normality plots and tests
- Continue
- Options – Check Exclude cases pairwise
- Continue



**Fig. 20.5** A Stem-and-Leaf pattern of hemoglobin levels

## SPSS Outputs [1]

Tests of Normality						
	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
Hemoglobin baseline	0.045	559	0.009	0.997	559	0.354

<sup>a</sup>Lilliefors significance correction

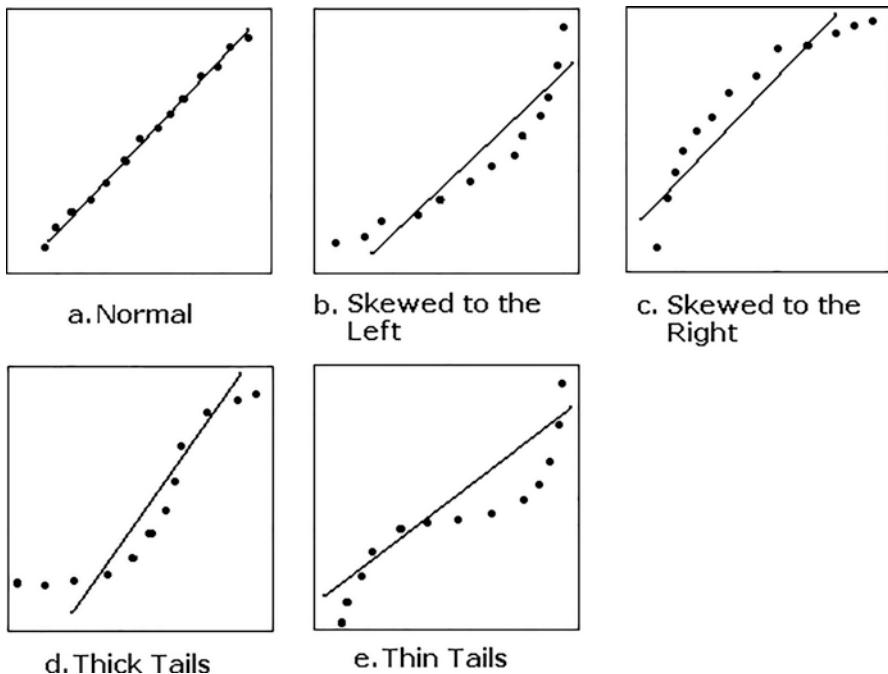
## Interpretation

The Shapiro-Wilk test shows that  $p$ -value = 0.354, indicating no significant departure from normality. In other words, the data are normally distributed.

## 3 What Tests to Do When: A Practical Guideline

### 3.1 Comparing Two Means and Independent Groups

- Distribution is normal
  - Population standard deviation is known, use  $z$ -test
  - Population standard deviation is not known, use the  $t$ -test

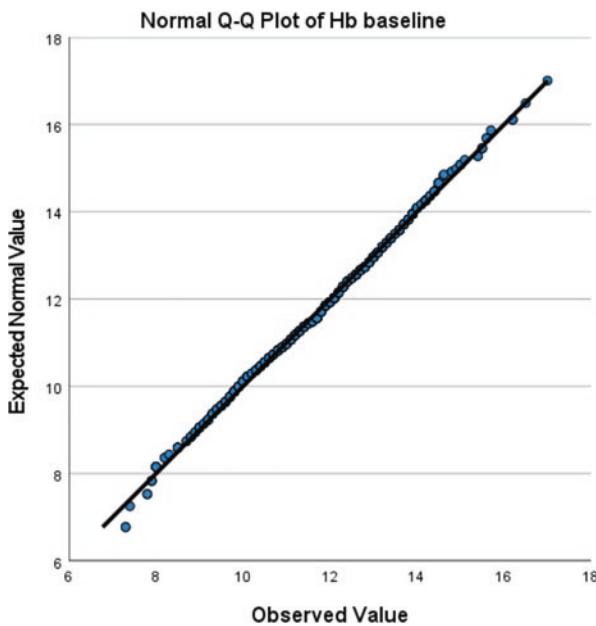


**Fig. 20.6** A Q–Q distribution showing normal (**a**), skewed distribution (**b**, **c**), thick (**d**) and thin (**e**) tailed plot

- Distribution not normal
  - Mann–Whitney  $U$  test

### 3.2 Comparing Two Means, Dependent Groups, or Paired Data

- You have data at two points for each individual
- Example: Morning and evening blood pressure; blood sugar measured before and after exercise
- Distribution is normal
  - Paired  $t$ -test
- Distribution nonnormal: nonparametric tests
  - Nominal scale – McNemar's test
  - Ordinal scale – Wilcoxon signed rank test



**Fig. 20.7** A Q–Q distribution shows normal distribution of hemoglobin values [1]

### 3.3 Compare Three or More Means

- Distribution is normal
  - ANOVA
- Distribution not normal
  - Kruskal–Wallis one-way ANOVA
  - Friedman's two-way ANOVA (multiple treatments)

### 3.4 Compare Two or More Qualitative (or Group) Variables

Example,

- Variable 1 = Education level (High school or less; Bachelor's degree; higher education) and
- Variable 2 = Knowledge about organ donation (good; poor)
- Chi-square test

### ***3.5 Association Between Two or More Variables***

- Distribution is normal – Pearson’s Correlation
- Distribution not normal – Spearman correlation

### ***3.6 Prediction of a Dependent Variable from Multiple Independent Variables***

- If the dependent variable is continuous – multiple linear regression
- If the dependent variable is dichotomous – multiple logistic regression

## **4 Data Analysis Addressing the Research Questions**

In the Iron study mentioned earlier, there were several research questions. This exercise will concentrate on two important research questions to analyze the data. These questions are as follows.

### ***4.1 Did the Hemoglobin Levels Improve After Treatment in Low-Income Postpartum Women?***

In this study, each woman had two data points: hemoglobin measured at baseline (hgb) and at follow-up (hbf). When your data values are paired measurements and the data distribution is normal, you use a paired *t*-test.

#### **Steps for SPSS**

- Go to Analyze
- Compare Means
- Paired Sample *t*-test
- Select Paired Variables: hgb (hemoglobin at baseline) and hbf (hemoglobin at follow-up)
- OK.

#### **SPSS Outputs [1]**

<b>Paired Samples Statistics</b>					
		Mean	<i>N</i>	Std. Deviation	Std. Error Mean
Pair 1	Hemoglobin baseline	11.685	168	1.7461	0.1347
	Hemoglobin at follow-up	12.320	168	1.2512	0.0965

Paired samples test								
			95% CI of mean difference					Significance
Mean difference	Std Deviation	Std Error Mean	Lower	Upper	t	df	One-sided p-value	Two-sided p-value
-0.635	1.7902	0.1381	-0.9072	-0.3618	-4.594	167	<0.001	<0.001

CI confidence interval

### Interpretations

Hemoglobin levels at follow-up (after treatment) significantly increased from the baseline levels (mean  $\pm$  SD),  $12.32 \pm 1.25$  vs.  $11.69 \pm 1.75$ ; 95% CI  $-0.91$  to  $-0.36$ ;  $p < 0.001$ .

## 4.2 Did the Women in the Three Treatment Groups Differ in Hemoglobin Status After Treatment?

The variable for the three treatment groups was “group.” The treatment groups were Control, Treatment Group I, and Treatment Group II.

The mean hemoglobin at follow-up was compared among the three groups by using one-way ANOVA and Tukey’s honestly significant difference (HSD) test (a post hoc test).

### Steps for SPSS

- Analyze
- Compare Means
- One-Way ANOVA
- Select Dependent List: hbf (hemoglobin at follow-up)
- PostHoc: Select Tukey
- Continue
- Options: Select Descriptive
- Continue
- OK

### SPSS Outputs

Descriptive								
	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum	
				Lower bound	Upper bound			
Control	12.592	1.0697	0.1327	12.327	12.857	9.5	14.9	

**Descriptive**

	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
				Lower bound	Upper bound		
Treatment 1	12.112	1.3110	0.1298	11.854	12.369	7.8	16.0
Treatment 2	<b>13.950</b>	0.3536	0.2500	10.773	17.127	13.7	14.2
Total	12.318	1.2476	0.0960	12.129	12.508	7.8	16.0

**ANOVA****Follow-up hemoglobin**

	Sum of Squares	df	Mean Square	F	Sig.
Between groups	14.556	2	7.278	4.893	0.009
Within groups	246.937	166	1.488		
Total	261.493	168			

**Multiple Comparisons****Dependent Variable: Follow-up hemoglobin****Tukey's HSD**

(I) treatment group	(J) treatment group	Mean difference (I - J)	Std. Error	Sig.	95% Confidence Interval	
					Lower bound	Upper bound
Control	Treatment 1	0.4805 <sup>a</sup>	0.1936	<b>0.037</b>	0.023	0.938
	Treatment 2	-1.3577	0.8756	0.270	-3.428	0.713
Treatment 1	Control	-0.4805 <sup>a</sup>	0.1936	0.037	-0.938	-0.023
	Treatment 2	-1.8382	0.8708	0.091	-3.898	0.221
Treatment 2	Control	1.3577	0.8756	0.270	-0.713	3.428
	Treatment 1	1.8382	0.8708	0.091	-0.221	3.898

<sup>a</sup>The mean difference is significant at the 0.05 level

**Interpretations**

- Based on the mean hemoglobin status at follow-up of the three groups (Descriptive Table), Treatment Group II (those who were given universal iron treatment, irrespective of hemoglobin status) had the highest level of hemoglobin (13.95) after treatment. Therefore, the universal treatment group gained the most benefit from iron treatment.
- The ANOVA table shows a *p*-value of 0.009. That means at least one of the three treatment groups is significantly different from another group regarding hemoglobin status at follow-up.

3. The Multiple Comparison Table shows the following results: Control versus Treatment I:  $p = 0.037$ , indicating that the mean difference in hemoglobin was higher among the Control group women than in Treatment I. There was no statistically significant difference in treatment among the other pairs.

## 5 Further Practice

1. Describe in your own words how the information about variables can help you decide the statistical test?
2. What type of statistical test would you do if the data are not normally distributed?
3. You conducted a study with 40 samples. You wanted to find the normality of your data by using a statistical test. What statistical test is most commonly used for the normality test?
4. Choose the correct answer:  
You have three variables, and all of them are categorical or grouped. What statistical test would you use to compare the variables?
  - (a)  $t$ -Test
  - (b) Chi-square test
  - (c) One-way ANOVA
  - (d) Correlation
  - (e) Regression
5. You have two variables, one is blood pressure measured as a continuous variable, and the second is gender (male and female). What statistical test would you use to compare the blood pressure between males and females?
  - (a) Independent  $t$ -test
  - (b) Chi-square test
  - (c) One-way ANOVA
  - (d) Dependent  $t$ -test
  - (e) Regression
6. You measured cholesterol at baseline and again after 6 months of taking Lipitor. What statistical test would you do to compare the cholesterol levels at baseline and after 6 months?
  - (a) Independent  $t$ -test
  - (b) Chi-square test
  - (c) One-way ANOVA
  - (d) Dependent  $t$ -test
  - (e) Regression

7. You measured the serum hemoglobin of 200 women; they were divided into three groups: young, middle-aged, and elderly. What statistical test would you use to compare the hemoglobin levels of the three groups?
  - (a) Independent *t*-test
  - (b) Chi-square test
  - (c) One-way ANOVA
  - (d) Dependent *t*-test
  - (e) Regression
8. You wanted to predict blood pressure from multiple independent variables such as age, gender, race, exercise, smoking, income, and stress level. Blood pressure is recorded as a continuous variable. What statistical test would you use to predict blood pressure from the independent variables?
  - (a) Chi-square test
  - (b) Multiple linear regression
  - (c) Multiple logistic regression
  - (d) ANOVA
  - (e) Correlation
9. You wanted to predict blood pressure from multiple independent variables such as age, gender, race, exercise, smoking, income, and stress level. Blood pressure is recorded as high blood pressure or normal blood pressure. What statistical test would you use to predict blood pressure from the independent variables?
  - (a) Chi-square test
  - (b) Multiple linear regression
  - (c) Multiple logistic regression
  - (d) ANOVA
  - (e) Correlation
10. You collected data from several variables such as blood pressure, age, gender, race, exercise, smoking, income, and stress level. What statistical test would you use to find the association between variables?
  - (a) Chi-square test
  - (b) Multiple linear regression
  - (c) Multiple logistic regression
  - (d) ANOVA
  - (e) Correlation

### Answer Keys

1. Information that are needed to decide a statistical test includes: (a) Hypothesis (if any); (b) research questions to address; (c) variable type; (d) scale of data; and (e) distribution of data (normal or non-normal)
2. Nonparametric test
3. Shapiro-Wilk test as the sample size is <50
4. (b)

5. (a)
6. (d)
7. (c)
8. (b)
9. (c)
10. (e)

## References

1. Mitra AK, Khoury A. Universal iron supplementation: a simple and effective strategy to reduce anaemia among low-income, postpartum women. *Public Health Nutr.* 2012;15(3):546–53. <https://doi.org/10.1017/S1368980011001261>. Available at: <https://www.cambridge.org/core/services/aop-cambridge-core/content/view/9F4DADAD18B159C234495CE363852B43/S1368980011001261a.pdf/div-class-title-universal-iron-supplementation-a-simple-and-effective-strategy-to-reduce-anaemia-among-low-income-postpartum-women-div.pdf>
2. Mishra P, Pandey CM, Singh U, Gupta A, Sahu C, Keshri A. Descriptive statistics and normality tests for statistical data. *Ann Card Anaesth.* 2019;22(1):67–72. [https://doi.org/10.4103/aca.ACA\\_157\\_18](https://doi.org/10.4103/aca.ACA_157_18).
3. Ghasemi A, Zahediasl S. Normality tests for statistical analysis: a guide for non-statisticians. *Int J Endocrinol Metab.* 2012;10(2):486–9. <https://doi.org/10.5812/ijem.3505>.
4. Curran-Everett D. Explorations in statistics: the assumption of normality. *Adv Physiol Educ.* 2017;41:449–53. <https://doi.org/10.1152/advan.00064.2017>.
5. Daniel WW, Cross CL. Biostatistics: a foundation for analysis in the health sciences. 10th ed. Hoboken: John Wiley & Sons; 2013.

# Chapter 21

## Data Analysis Using SPSS: Jackson Heart Study



Clifton C. Addison and Brenda W. Campbell Jenkins

### Learning Objectives

After completing this chapter, you will be able to:

- Recognize and explain the statistical procedures required to describe the characteristics of interest
- Write effective hypotheses and research questions to conduct appropriate statistical tests
- Recognize and explain statistical procedures required to test the hypotheses of difference and association
- Demonstrate an understanding of and the ability to make statistical inferences after hypothesis testing
- Interpret and draw conclusions from statistical analysis outputs in SPSS

## 1 Introduction

In this chapter, descriptive statistics are used to examine the characteristics of the Jackson Heart Study (JHS) participants. Categorical variables used to describe the participants include the following: county of residence, age, gender, socioeconomic status (measured by educational level), and income. Descriptive statistics used to examine the categorical variables that describe the characteristics of the JHS cohort include frequencies and percentages. Continuous variables used to describe the participants include the following: age, body mass index (BMI), fasting glucose, total cholesterol, alcohol drinking, and smoking (pack years). Descriptive statistics used to examine the continuous variables that describe the characteristics of the JHS

---

C. C. Addison (✉)

Department of Epidemiology and Biostatistics, School of Public Health, College of Health Sciences, Jackson State University, Jackson, MS, USA  
e-mail: [clifton.addison@jsums.edu](mailto:clifton.addison@jsums.edu)

B. W. C. Jenkins

Jackson Heart Study, Jackson State University, Jackson, MS, USA  
e-mail: [brenda.w.campbell@jsums.edu](mailto:brenda.w.campbell@jsums.edu)

cohort include measures of central tendency (e.g., mean, median, and mode) and measures of variability (standard deviation, variance, and range).

Inferential statistics are also used to examine the JHS cohort and test hypotheses of interests. The inferential tests performed involve parametric statistics. The parametric tests performed include the independent samples *t*-test to address differences between two group means on the outcome of interest, the one-way analysis of variance (ANOVA) to address differences between three or more group means on the outcome of interest, and the Pearson's Rho correlation test to examine the bivariate relationship between continuous variables.

## 2 The Jackson Heart Study

The Jackson Heart Study (JHS), initiated in 1998 and funded by the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD), is a longitudinal investigation of genetic and environmental risk factors associated with the disproportionate burden of cardiovascular disease in African Americans. The JHS exemplifies a unique collaborative model among three institutional partners, Jackson State University (JSU), Tougaloo College, and the University of Mississippi Medical Center (UMMC), the Jackson community, and the National Institutes of Health to discover and test best practices for eliminating health disparities. The mission of the JHS is to elucidate the reasons for the greater prevalence of cardiovascular disease among African Americans and to uncover new approaches for reducing this health disparity. The vision of the JHS is to transform a history of African Americans' heart disease into a legacy of heart health through research and by translating and disseminating research results.

### 2.1 Variables Used in the Analyses

The JHS recruited 5306 African American residents living in the Jackson, Mississippi, metropolitan area of Hinds, Madison, and Rankin Counties. JHS participants received three back-to-back clinical examinations (Exam 1 in 2000–2004, Exam 2 in 2005–2008, and Exam 3 in 2009–2013) that have generated extensive longitudinal data on traditional and putative cardiovascular disease risk factors and measures of subclinical cardiovascular disease from echocardiography, cardiac magnetic resonance imaging, and computed tomography scans. Biological samples (i.e., blood and urine) have been assayed for putative biochemical risk factors and stored for future research. DNA has been extracted, and lymphocytes have been cryopreserved for studies of candidate genes, genome-wide scanning, expression, and other –omics investigations. The Mississippi State Department of Health was added to the group of JHS partners in 2018, and Exam 4 was scheduled to begin in

2020. It includes a clinical examination and investigation of the link between cardiovascular health and brain health [1].

The JHS variables used in the analyses to test the hypotheses in this chapter include the following:

1. **Independent *t*-test:** Independent Variable = Gender; Dependent (Outcome) Variable = BMI Level
2. **ANOVA:** Independent variable = BMI Group; Dependent (Outcome) Variable = Fasting Glucose
3. **Pearson Rho correlation test:** Variables = Fasting Glucose, BMI

### 3 Practical Application of the Analytic Plan Using Jackson Heart Study Data

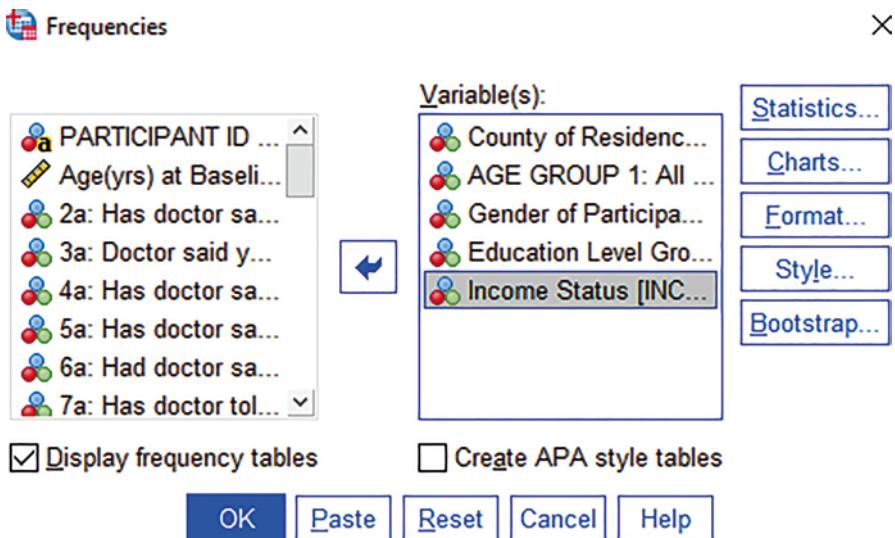
#### 3.1 Describing Characteristics of Groups Using SPSS [2]

Before you start data analysis, the first step is to make sure that the data are clean. There are procedures for cleaning the data. Second, you must have thorough knowledge about the variables. Third, you should have some hypotheses and/or research questions. When you open the data spreadsheet, go to the top row of the SPSS data editor, and you will find the menu bar where the commands are located. Review each menu, then click on the appropriate command for your chosen procedure.

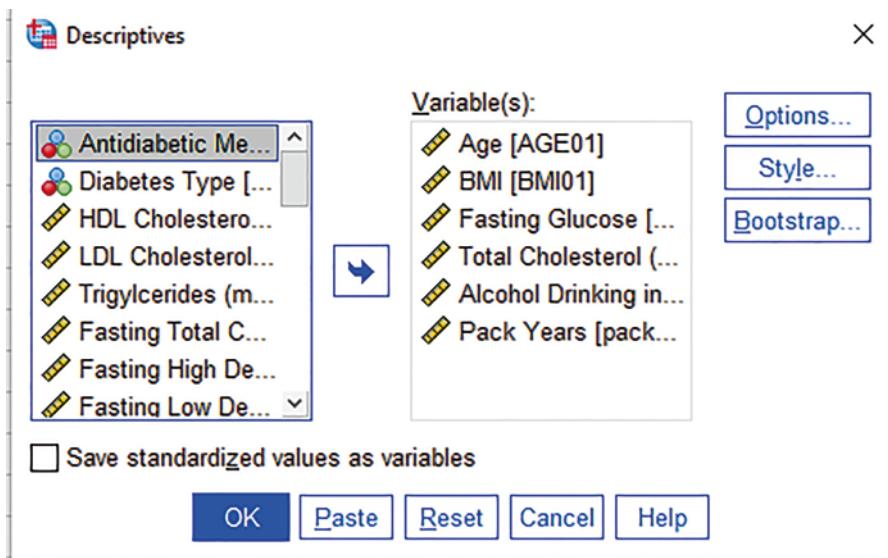
To describe a variable of interest that is a category (a variable with more than one level), select “Frequencies” as the appropriate analytic descriptor. To do that, click on:

**Analyze > Descriptive > Frequencies** – Highlight the variable(s) of interest, then click the arrow that will insert them into the variable box in preparation for conducting the analysis. Once the required variables are entered, click OK. See the box below.

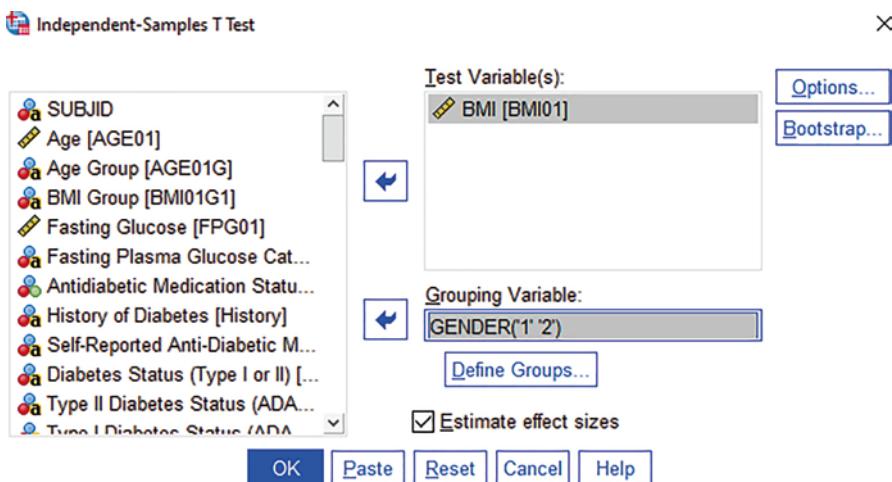
On the next few pages, we display five figures (Figs. 21.1, 21.2, 21.3, 21.4, and 21.5) to illustrate steps of some basic statistical analysis using SPSS. The figures are generated when variables from JHS data are extracted by the variable finder for analysis using SPSS. The authors have permission to use JHS data for publication in this book. Each figure is a screenshot obtained from SPSS analysis done in our computer lab using the JHS data. Please note that the images used to describe the



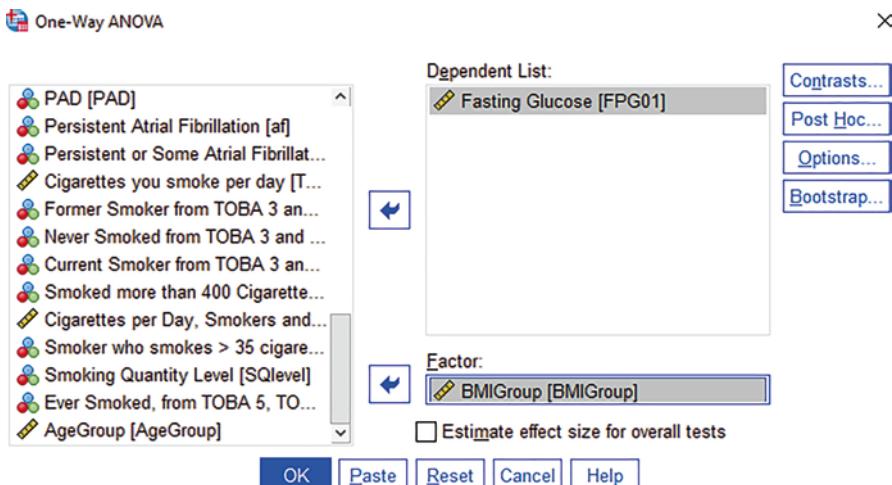
**Fig. 21.1** Steps showing the frequency distribution analysis. (Source: This figure is a screenshot of the SPSS data analysis using JHS data at our lab [2, 3]. This is generated when variables are entered as an initial step in analyzing the frequency distribution)



**Fig. 21.2** Steps showing the descriptive data analysis. (Source: This figure is a screenshot of the SPSS data analysis using JHS data at our lab [2, 3]. This is generated when variables are entered as an initial step of analyzing descriptive statistics)



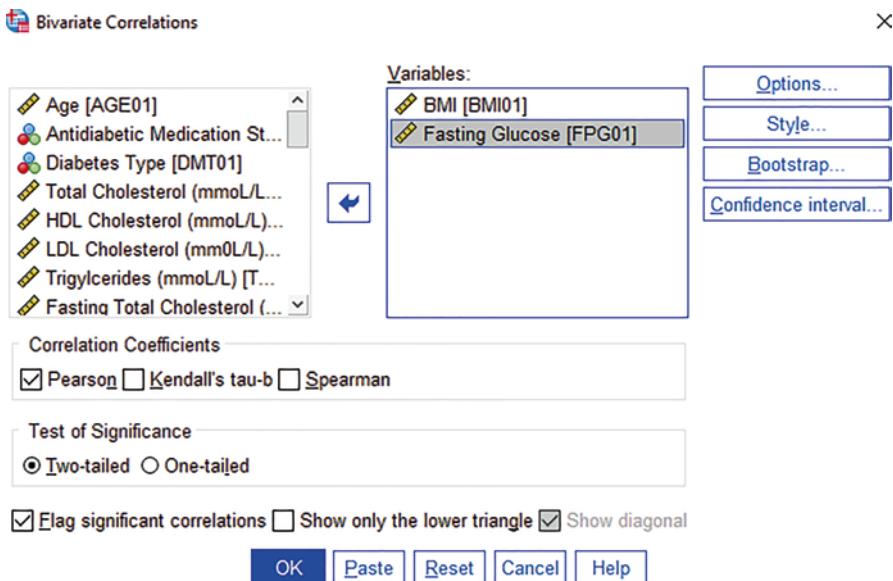
**Fig. 21.3** Steps showing the independent samples  $t$ -test. (Source: This figure is a screenshot of the SPSS data analysis for independent sample  $t$ -test with JHS data [2, 3]. This is generated when variable BMI is entered as a test variable and GENDER is used as a Grouping Variable for the test. This figure is a screen print of the SPSS page that is generated when variable (BMI) is entered for an independent sample  $t$ -test with GENDER)



**Fig. 21.4** Steps showing the one-way analysis of variance (ANOVA) test. (Source: This figure is a screenshot of the SPSS data analysis for one-way ANOVA test using JHS data [2, 3]. Here, fasting glucose is a dependent variable and BMIGROUP is an independent factor used for the analysis)

analytic processes in this chapter do not necessarily include all of the variables in the dataset because of the large size of the JHS datasets.

The following is the “Frequencies” SPSS output that enables the researcher to describe some of the characteristics of the sample with regards to group membership:



**Fig. 21.5** Steps showing the bivariate Pearson correlation test. (Source: This figure is a screenshot of the SPSS data analysis for bivariate correlation test using JHS data [2, 3]. This is generated when variables such as BMI and fasting glucose are entered for the analysis)

### 3.1.1 Frequency Tables

The SPSS “Statistics” table below provides general information about the selected variables, for example, the number of participants with data in each category as well as the number of participants with missing data. The tables that follow represent the distribution of the sample according to each of the levels of each variable examined. In reporting the findings, the researcher can then indicate or describe how many or what percentage of the participants belong to each level (group) of the respective variables.

#### *SPSS Frequency Analysis*

The frequency procedure produces two tables: the Statistics Table and the Frequency Table. The Statistics Table provides an account of the total number of participants who provided information for each variable selected for examination, as well as the number of participants from the sample who did not provide any data (missing). The Frequency Table provides the frequency (number) and percentage of occurrence for each level of the examined variable. For variable #1, County of Residence, the table indicates that 2472 of the participants, representing 84.6% of the total number of participants, resided in Hinds County, 244 of the participants, representing 8.4% of the participants, resided in Madison County, and 107 of the participants, representing 3.7% of the participants, resided in Rankin County. There were 99 participants who did not provide any information for this question (missing data).

## ► Frequencies

### Statistics

	County of Residence	AGE GROUP 1: All Participants (Entire Study Population)	Gender of Participant	Education Level Group 2	Income Status
N	Valid	2823	2922	2922	2912
	Missing	99	0	0	10

### Frequency Table

#### County of Residence

	N	%
Hinds	2472	84.6%
Madison	244	8.4%
Rankin	107	3.7%
Missing System	99	3.4%

#### AGE GROUP 1: All Participants (Entire Study Population)

	N	%
21-34	36	1.2%
35-44	340	11.6%
45-54	620	21.2%
55-64	949	32.5%
65-74	741	25.4%
75-84	226	7.7%
85 and above	10	0.3%

#### Gender of Participant

	N	%
Female	1958	67.0%
Male	964	33.0%

### **Education Level Group 2**

	N	%
Less than High School	690	23.6%
High School/GED	630	21.6%
Beyond High School, Less than Bachelors	749	25.6%
Bachelors Degree or Higher	843	28.9%
Missing System	10	0.3%

### **Income Status**

	N	%
Low	432	14.8%
Lower-Middle	691	23.6%
Upper-Middle	690	23.6%
Affluent	650	22.2%
Missing System	459	15.7%

To describe a variable of interest, when the variable of interest is a number (a quantitative measurement that represents/describes the individual), select “Descriptives,” click on:

**Analyze > Descriptives > Descriptives** – Highlight the variables of interest and then click the arrow that will take them into the **Variable Box**. Once the required variables are entered, click OK.

The following is the “Descriptives” SPSS output that enables the researcher to describe some of the characteristics of the sample with regards to group performance:

#### **SPSS Descriptive Analysis**

*The “Descriptive Statistics” table below provides general information about the selected variables, like the number of participants providing measures, the minimum score, the maximum score, the mean, and the standard deviation for each individual variable. The table that follows represents the quantitative description of each variable examined. In reporting the findings, you can describe the average measure that represents the overall group performance (mean), as well as the average difference of individual measures from the mean (the standard deviation). In the Descriptive Statistics table below, the minimum age of the participants is 20 years old, and the maximum age is 95 years old. The mean age is 54.87 years old, and the standard deviation is 12.848.*

## ► Descriptives

**Descriptive Statistics**

	N	Minimum	Maximum	Mean	Std. Deviation
Age	5301	20	95	54.87	12.848
BMI	5292	14.6296296	91.7972624	31.75191114	7.241175882
Fasting Glucose	4830	36	503	99.98	32.981
Total Cholesterol (mmoL/L)	5170	1.69	12.05	5.0979	1.03042
Alcohol Drinking in the Past 12 months (days/weeks)	2329	0	7	1.20	1.760
Pack Years	5152	0	228	7.26	16.673
Valid N (listwise)	2071				

## 3.2 Testing Hypothesis of Difference [3]

### 3.2.1 Testing Differences Between Two Groups When the Dependent Variable (Outcome) Is a Continuous Variable

Use the independent samples *t*-test if the intention is to test for differences between **two groups** (two samples), which are independent. In this dataset, “gender” is a group variable, with male and female being two independent groups.

**Hypothesis 1 ( $H_0$ ):** There is no significant difference between males and females students in their BMI levels.

**Alternative Hypothesis ( $H_a$ ):** There is a significant difference between male and female in their BMI levels.

Since the variable gender has two levels, male and female, the appropriate test of difference is the independent sample *t*-test.

The independent sample *t*-test is the appropriate test to compare the means when the groups are independent and there are only two groups. Your research hypothesis (which is also the alternative hypothesis) is to find statistical evidence that the two sample means are significantly different. The independent samples *t*-test, a parametric test, is the most commonly used test statistics when the distribution of the variable is normal. For a nonnormal distribution, you will choose an equivalent nonparametric test. The equivalent nonparametric test for an independent sample *t*-test is called the Mann–Whitney *U* test. Therefore, before selecting a statistical test, you must find the distribution of the data.

The independent *t*-test has several different names. Depending on the textbook you are using, it may be called (1) independent *t*-test, (2) independent measures *t*-test, (3) independent two-sample *t*-test, (4) Student’s *t*-test, and (5) two-sample *t*-test. So, try not to let these different terms confuse you when you find them.

To conduct the independent sample *t*-test, you must be able to clearly identify the independent variable, sometimes called the grouping variable, the testing variable, or the outcome variable.

### ***Assumptions for t-Test***

The data must meet certain assumptions in order to conduct the independent *t*-test. Here are the assumptions:

1. Assumption of independence: the two groups must be mutually exclusive, meaning that if you belong to one group, you cannot belong to the other.
2. Assumption of random selection: sample should be randomly selected.
3. Assumption of normality of the dependent variable for each group: the distribution must be normal.
4. Assumption of homogeneity: there must be equal variances between the groups, that is, variability should be similar.

The null hypothesis and the alternative can be expressed in two different ways:

$H_0: M_1 = M_2$  (the mean of Group 1 is equal to the mean of Group 2 – the two means are equal)

$H_a: M_1 \neq M_2$  (the mean of Group 1 is not equal to the mean of Group 2 – the two means are not equal), or

$H_0: M_1 - M_2 = 0$  (the difference between the two means is equal to 0- that is the same thing as saying that the two means are equal)

$H_a: M_1 - M_2 \neq 0$  (the difference between the two means is not equal to 0 – that is the same thing as saying that the two means are not equal)

To compute to the independent *t*-test, go to the SPSS Data Editor:

Click **Analyze > Compare Means > Independent-Samples t-Test** – The independent-samples *t*-test window opens, where you can specify the variables you want to use in the analysis. All of the variables in your dataset can be found in the list of variables on the left side. Highlight the test variable(s) (dependent variable/outcome), then click the arrow that will take it into the **Test Variable Box**. Then, highlight the independent variable (gender), click the arrow, and send it to the **Grouping Variable** box. The “**Define Groups**” command will become active. You must define the range of categories of your grouping variable (e.g., *I–2, 0–1*). In case you do not remember the codes used for male and female, click on “Utilities” at the command row (which is at the top line). Then click on the variable “Gender,” it will show the code (such as male = 1, female = 2, or other codes). Once you know the codes, click on the **Define Groups** box and put the numbers that correspond to the two groups that are being tested (in the **Group 1** box, insert the code 1 that represents male; in the **Group 2** box, insert the code 2 that represents female). Next, click “**Continue**,” and then “**OK**.” You will find the data output for the results.

The result of the *t*-test analysis is as follows: Two tables are generated in the SPSS output: **Group Statistics** and **Independent Samples Test**.

### **SPSS Independent Sample t-Test**

The first section, **Group Statistics**, provides basic information about the group comparisons, including the sample size ( $n$ ), mean, standard deviation, and standard error for BMI by gender. In this example, there are 3355 female participants and 1937 male participants. The mean BMI for female participants is 32.8, and the mean BMI for male participants is 29.9.

The second section, **Independent Samples Test**, displays the results most relevant to the independent samples  $t$ -test. There are two parts that provide different pieces of information: (A) Levene's test for equality of variances and (B)  $t$ -test for equality of means that provides the calculated significance value ( $p$ -value).

This section provides the evidence for you to form your conclusions about whether the difference between the two means is statistically significant. To do this, you must remember *the decision rule*. *The decision rule* is the rule you must follow when you want to determine whether to accept or fail to reject the null hypothesis ( $H_0$ ), or whether you will reject  $H_0$ . Here is *the decision rule*: If your calculated significance value ( $p$ -value) that was derived from your analysis is  $\leq 0.05$  (the critical comparison value of alpha), you reject the null hypothesis and come to the conclusion that there is a statistically significant difference. You will have to conclude that there is a significant difference between the two means: **Group 1** (Females) and **Group 2** (males). If your calculated significance value ( $p$ -value) is greater than 0.05 (e.g., 0.051), you will conclude that there is no significant difference between the two groups. In this case, you will accept the null hypothesis ( $H_0$ ) that there is no significant difference between the two means: **Group 1** (females) and **Group 2** (males). The significance level computed in the table is  $p < 0.001$ . Since that value is less than 0.05, the decision is to reject the null hypothesis and mention that the difference between male and female participants in their BMI is statistically significantly different.

In the data output, you will find 95% confidence intervals. In the table below ( $t$ -test), 95% confidence intervals of the difference between two means are 2.58–3.38. These intervals show that the mean difference in the population lies between these two numbers is 2.58 and 3.38.

### ***t*-Test**

<b>Group Statistics</b>					
Gender		<i>N</i>	Mean	Std. Deviation	Std. Error Mean
BMI	Female	3355	32.84387376	7.599805992	.1312067698
	Male	1937	29.86056649	6.131522448	.1393168070

## Independent Samples Test

Levene's Test of Equality of variances				<i>t</i> -test for Equality of Means				95% Confidence interval of Diff.		
	<i>F</i>	Sig.	<i>t</i>	<i>df</i>	Significance One-sided <i>p</i>	Two-sided <i>p</i>	Mean Diff.	Std Error diff.	Lower	Upper
BMI	102.169	<.001	14.729	5290	<.001	<.001	2.983307270	.2025449327	2.586235646	3.380378894
Equal variances assumed										
Equal variances not assumed		15.589	4740.643	<.001	<.001	2.983307270	.1913749961	2.608123380	3.358491160	

## Independent Samples Effect Sizes

BMI	Standardized <sup>a</sup>	Point estimate	95% Confidence Interval
Cohen's d	7.097780054	.420	.364 .477
Hedges' correction	7.098786554	.420	.364 .477
Glass' delta	6.131522448	.487	.429 .544

Cohen's d uses the pooled standard deviation

Hedges' correction uses the pooled standard deviation, plus a correction factor

Glass's delta uses the sample standard deviation of the control group

<sup>a</sup>The denominator used in estimating the effect size

### 3.2.2 Differences Between Three or More Groups When the Dependent Variable (Outcome) Is a Number (or Continuous Variable)

If the intent is to compare the means of more than two groups and the data are normally distributed, the appropriate test is the one-way analysis of variance (ANOVA).

**Hypothesis 2 ( $H_0$ )** There is no significant difference between the BMI levels of students based on their systolic blood pressure category. Since the variable systolic blood pressure category has four levels (Normal, Prehypertensive, Stage 1 Hypertension, and Stage 2 Hypertension), the appropriate test of difference is a one-way ANOVA.

To compute the ANOVA test, go to the **SPSS Data Editor**:

Click **Analyze > Compare Means > One-Way ANOVA** – Highlight the test variable(s) (dependent List), then click the arrow that will take it into the **Dependent List Box**. Then, highlight the independent variable (Systolic BP category), click the arrow, and send it to the **Factor** box. Next, click “**OK**.” The computer will conduct your analysis.

The result of the one-way ANOVA test analysis is as follows:

#### **SPSS ANOVA Analysis**

The data in the ANOVA table indicate that there is a significant difference in the fasting glucose level of the participants based on age ( $p < 0.001$ ). Since the ANOVA test revealed a significant difference, then that information warrants the performance of a post hoc test. The post hoc test is a multiple comparison test that enables the researcher to determine exactly where the differences exist when multiple group levels of a variable are compared. In this case, the post hoc test called Tukey’s honestly significant difference (HSD) was performed.

► **Oneway**

**ANOVA**

**Fasting Glucose**

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	111709.968	2	55854.984	52.405	<.001
Within Groups	5137337.898	4820	1065.838		
Total	5249047.866	4822			

**Post Hoc Tests**

**Multiple Comparisons**

Dependent Variable: Fasting Glucose

Tukey HSD

(I) BMI Group	(J) BMI Group	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
Less than 25	25-30	-7.634*	1.480	<.001	-11.10	-4.17
	30 and above	-13.592*	1.389	<.001	-16.85	-10.34
25-30	Less than 25	7.634*	1.480	<.001	4.17	11.10
	30 and above	-5.958*	1.048	<.001	-8.41	-3.50
30 and above	Less than 25	13.592*	1.389	<.001	10.34	16.85
	25-30	5.958*	1.048	<.001	3.50	8.41

\*. The mean difference is significant at the 0.05 level.

### 3.3 Testing Hypothesis of Association (Relationship/Correlation) [3]

**Hypothesis 3 ( $H_{03}$ )** There is no significant association between the BMI levels of students and their systolic blood pressure. Since the variables “BMI” and “Systolic Blood Pressure” are both quantitative numbers, the appropriate test of association is the Pearson correlation ( $r$ ).

To compute Pearson correlation test, go to the **SPSS Data Editor**:

Click **Analyze > Correlate > Bivariate** – Highlight the two variables, “BMI” and “Systolic Blood Pressure,” then click the arrow that will take them into the **Variables** Box. Make sure that the **Pearson** box under “Correlation Coefficient” is checked. Then, Click OK. The computer will generate the results.

Below is the Pearson correlation test output:

As seen in the “Correlations” table below, BMI is significantly correlated with Fasting Glucose ( $p \leq 0.001$ ). You should analyze several things for a correlation

test: (1) **Strength of association:** the value of “ $r$ ” (also called the correlation coefficient) will tell you the strength of the association. A perfect correlation is 1. Any number close to 1 is very strong. In general, if the  $r$ -value is greater than 0.5, it is considered high. In this case,  $r = 0.138$ ; if it is less than 0.5, which means the association is weak. (2) **Direction of the association:** If the value of “ $r$ ” is positive, the association between the variables is direct, and if the value of “ $r$ ” is negative, the relationship is inverse or indirect. In this case, the value of “ $r$ ” is positive. Therefore, the relationship between BMI and Fasting Glucose is direct, meaning a higher value of BMI is related to a higher level of Fasting Glucose. (3) **Statistical significance:** Since the  $p$ -value  $<0.001$ , the association between the two variables is statistically significant.

To explain the correlation between two variables further, you should calculate  $r^2$  (square the value of “ $r$ ”). In this case,  $r^2 = (0.138)^2$ .

The calculated value of  $r^2 = 0.019$ . It is called the **coefficient of determination**. Based on this number, only 0.019% of the variation of Fasting Glucose (the dependent variable) can be determined by BMI. You must make a judgment – statistical significance tells you that the data are significant; however, it is also important to know how much variability of the dependent variable can be assessed by the association of an independent variable. Although a  $p$ -value may be significant, if the coefficient of determination is low, the observed association may not be very relevant to your study objectives.

### **SPSS Pearson Correlation Analysis**

#### ➔ **Correlations**

**Correlations**

		BMI	Fasting Glucose
BMI	Pearson Correlation	1	.138**
	Sig. (2-tailed)		<.001
	N	5292	4823
Fasting Glucose	Pearson Correlation	.138**	1
	Sig. (2-tailed)	<.001	
	N	4823	4830

\*\*. Correlation is significant at the 0.01 level (2-tailed).

If the relationship between two variables is significant, then additional details should be provided about the strength of the relationship (low/weak, medium, or high, depending on how close it is to “1” or “0”), as well as the direction of the relationship (negative or positive).

## 4 Recap and Review

**Inferential statistics** involves two types of analysis: (1) hypothesis testing and (2) confidence intervals.

**Hypothesis testing** involves using a statistical analysis to enable the researcher to address the question(s) that precipitated the research in the first place. When one conducts research, the initial goal is to answer a particular question(s) about a subject of interest (problem of the study). You start off with a statement of the problem (problem statement), and a few research questions and research hypotheses.

## 5 How to Write the Null Hypothesis [4, 5]

In the case of a hypothesis of difference:

1. There is no significant difference between Group A and Group B on variable  $Y$

Symbols used for writing the null hypothesis

2.  $H_0: M_1 = M_2$ , (when the interest is with sample means)
3.  $H_0: \mu_1 = \mu_2$ , (when the interest is with population means)
4.  $H_0: M_1 - M_2 = 0$
5.  $H_0: \mu_1 - \mu_2 = 0$

If the results of the analyses demonstrate that items 2–5 are true, we accept the null hypothesis ( $H_0$ ) of the statement as it is framed in item 1 above.

Note: 1–5 are all saying the same thing, even though the narrative is different.

In the case of a Hypothesis of Association (Relationship or Correlation):

1. There is no significant association between variable  $X$  and variable  $Y$ . The appropriate test is the Pearson  $r$  (for a normal distribution) or Spearman Rho (for a nonnormal distribution), depending on the distribution of the data examined.

Symbols used for writing the null hypothesis when testing the hypothesis of association:

2.  $H_0: r = 0$ , where  $r$  is the correlation coefficient.

If  $r = 0$ , the decision would be to accept the null hypothesis ( $H_0$ ).

Under the hypothesis of association, one may also compute a multivariate analysis, such as a regression analysis. Multivariate analysis is done to predict any significant relationship between a dependent variable and independent variables. Multivariate analysis is discussed in detail in Chap. 22.

## 6 Further Practice

1. Review the section on frequency analysis and describe the frequency distribution of the variables: age group, gender, education level, and income status. Use the narrative for the variable county of residence as an example.
2. Review the data in the Descriptive Statistics table and describe the distribution of the variables: BMI, Fasting Glucose, Total Cholesterol, Alcohol Drinking, and Pack years (Cigarette Smoking). Use the narrative for the variable County of Residence as an example.
3. Develop hypotheses for conducting the independent *t*-test using variables located in the variable box for the independent samples *t*-test. See Figure 21.3 for the variable list.
4. Develop null and alternative hypotheses for conducting the one-way ANOVA using variables located in the variable box for the one-way ANOVA. See Figure 21.4 for the variable list.
5. Examine the post hoc test (Tukey's HSD) and report about where the differences occurred based on age.  
Hint: Examine the *p*-value (significance level) in the table to make that determination.
6. Develop null and alternative hypotheses for conducting the Pearson correlation test using variables located in the variable box for bivariate correlations. See Figure 21.5 for the variable list.

## References

1. About the Jackson Heart Study. Available at: <https://www.jacksonheartstudy.org/About/About-The-JHS>. Accessed 2 Aug 2022.
2. Addison CC, Jenkins BW, Sarpong D, Wilson G, Champion C, Sims J, White MS. Relationship between medication use and cardiovascular disease health outcomes in the Jackson Heart Study. *Int J Environ Res Public Health*. 2011;8(6):2505–15.
3. Jenkins BWC, Addison C, Wilson G, Liu J, Fortune M, Robinson K, White M, Sarpong D. Association of the joint effect of menopause and hormone replacement therapy and cancer in African American women: the Jackson Heart Study. *Int J Environ Res Public Health*. 2011;8:2491–504. <https://doi.org/10.3390/ijerph8062491>.
4. Frost J. Hypothesis testing—an intuitive guide for making data drive decisions. 1st ed. Statistics by Jim Publishing; 2020. ISBN-13: 978-1735431154; ISBN-10: 173543115X.
5. Gertsmann BB. Basic biostatistics: statistics for public health practice. 2nd ed. Sudbury: Bartlett and Jones; 2015. ISBN: 978-1-284-03601-5.

# Chapter 22

## Multiple Linear Regression and Logistic Regression Analysis Using SAS



Azad R. Bhuiyan and Lei Zhang

### Learning Objectives

After completing this chapter, you will be able to:

- Understand concepts of regression analyses
- Apply SAS code for various regression analyses
- Describe the Bogalusa Heart Study data
- Use model fit assumptions
- Assess multicollinearity
- Apply trend analysis
- Interpret SAS results of multiple regression

## 1 Introduction

Regression analysis describes and quantifies the relationship between dependent and independent variables. The term regression indicates the estimation or prediction of the average value of one dependent variable for a given independent variable [1, 2]. There are numerous types of regression models used in data science. This chapter covers simple and multiple linear regression and logistic regression, which are widely used in public health research [1, 2]. For example, in a study, our group wanted to quantify the relationship between birth weight and high-sensitivity C-reactive protein (hs-CRP) [3]. Regression analyses are most appropriate to find out the relationship. In that study, our dependent variable was hs-CRP (a continuous variable), and the independent variables included multiple variables, such as race, sex, age, body mass index (BMI), and current smoking in the model [3]. In this chapter, we will use our study data, which is part of a larger study known as the

---

A. R. Bhuiyan (✉)

Department of Epidemiology and Biostatistics, School of Public Health, College of Health Sciences, Jackson State University, Jackson, MS, USA  
e-mail: [azad.r.bhuiyan@jsums.edu](mailto:azad.r.bhuiyan@jsums.edu)

L. Zhang

School of Nursing, University of Mississippi Medical Center, Jackson, MS, USA  
e-mail: [lzhang2@umc.edu](mailto:lzhang2@umc.edu)

Bogalusa Heart Study (BHS). Gerald S. Berenson, a Bogalusa native and a cardiologist, founded the BHS, a community-based study on the early natural history of cardiovascular diseases funded by the National Heart, Lung, and Blood Institute, in 1972 [4]. The study cohort consisted of 776 black and white participants (28% black, 43% male), aged 24–43 years (mean 36.1 years).

## 1.1 Linear Regression

In this case, the dependent variable is continuous, and the independent variable is either continuous or categorical. Types of linear regression include:

- **Simple linear regression:** The model comprises one continuous dependent variable and one independent variable. In the above example, hs-CRP is a continuous independent variable, and birth weight is an independent variable. The regression equation for simple linear regression is  $y = b_0 + b_1$  (birth weight).
- **Multiple linear regression:** The model comprises one continuous dependent variable and more than one independent variable. Suppose your research interest is to investigate low birth weight with hs-CRP. The dependent variable is hs-CRP, and the independent variables are birth weight, race, sex, BMI, and cigarette smoking. The regression equation is:  $y$  (hs – CRP) =  $b_0 + b_1$ (birthweight) +  $b_2$ (race) +  $b_3$ (sex) +  $b_4$ (BMI) +  $b_5$ (smoking) +  $\varepsilon$ .

## 1.2 Assumptions for Linear Regression

To qualify for a regression analysis, the data must meet the following assumption [5, 6]:

1. The sample must be representative of the population.
2. The dependent variable should be normally distributed.
3. The relationship between the dependent and independent variables must be linear.
4. The dependent variable's distribution must have equal variability for any independent variable value. This refers to a condition called homoscedasticity. Homoscedasticity in a model means that the error is constant along the values of the dependent variable. The best way to check homoscedasticity is to make a scatterplot with the residuals against the dependent variable. The regression equation uses an error term to accommodate the variability. Therefore,  $y = \beta_0 + \beta_1(x) + \varepsilon$ .
5. In multiple regression, a high correlation between the independent variables cannot establish the true effect of the independent variables on the dependent variable. This issue of high correlation between multiple independent variables in a regression analysis is called multicollinearity. It is a common assumption that you test before selecting the variables for the regression model.

### 1.3 Model Fit Assumption Before Applying a Regression Model

#### 1.3.1 Test of Normality: Dependent Variable TR\_CRP (True CRP)

SAS code and the data are as follows [7]:

```
proc univariate plot normal;
var TR_CRP;
histogram/normal;
run;
```

Proc univariate provides a wider range of basic statistics, with the option for probability plots, normal for requesting tests for normality, the histogram for histogram, and option normal for bell-shaped curved on the histogram. SAS output shows that the mean and median values of True C-reactive protein (TR\_CRP) differ. The Skewness and Kurtosis values are not zero. Therefore, CRP is not normally distributed. Test for normality provides four tests for normality from proc univariate procedure. Null hypothesis ( $H_0$ ): data are normally distributed; alternate hypothesis ( $H_a$ ): not normally distributed (Table 22.1). For example, the Shapiro–Wilk test provides  $p < 0.0001$ . Therefore, we reject the null hypothesis of normality and accept that TR\_CRP is not normally distributed. Normal quantiles plot shows that it is an S-shaped curve with heavy-tailed distributions, and the histogram also shows that the data are skewed to the right (Fig. 22.1).

#### SAS code for influential and collinearity diagnosis [7]

SAS code in model option: *r* provides residual analysis, *influence* for influence statistics, *vif* for variance influence factor statistics, and *collinoint* for collinearity diagnosis with intercept adjustment.

```
proc reg;
model TR_CRP=bwkg1 race sex age bmi cursmk/r influence vif collinoint;
label TR_CRP= "True value of c-reactive protein"
      bwkg1= "Birth weight in kg"
      race= "Black vs. White"
      sex = "Female vs. Male"
      age = "Age in years"
      bmi = "Body mass index"
      cursmk= "Currently smoking";
run;
```

**Table 22.1** Univariate analysis of dependent variable

The UNIVARIATE Procedure				
Variable: TR_CRP (True value of c-reactive protein)				
Moments				
<b>N</b>	776	<b>Sum Weights</b>	776	
<b>Mean</b>	2.77324742	<b>Sum Observations</b>		2152.04
<b>Std Deviation</b>	3.21678261	<b>Variance</b>		10.3476903
<b>Skewness</b>	1.83978885	<b>Kurtosis</b>		2.94212516
<b>Uncorrected SS</b>	13987.5994	<b>Corrected SS</b>		8019.46002
<b>Coeff Variation</b>	115.993351	<b>Std Error Mean</b>		0.11547577
Basic Statistical Measures				
Location		Variability		
<b>Mean</b>	2.773247	<b>Std Deviation</b>		3.21678
<b>Median</b>	1.520000	<b>Variance</b>		10.34769
<b>Mode</b>	0.410000	<b>Range</b>		14.97000
		<b>Interquartile Range</b>		2.98500
Tests for Normality				
Test		Statistic		p Value
Shapiro-Wilk		W	0.759128	Pr < W
Kolmogorov-Smirnov		D	0.196026	Pr > D
Cramer-von Mises		W-Sq	11.22603	Pr > W-Sq
Anderson-Darling		A-Sq	62.7758	Pr > A-Sq
<0.0001				
<0.0100				
<0.0050				
<0.0050				

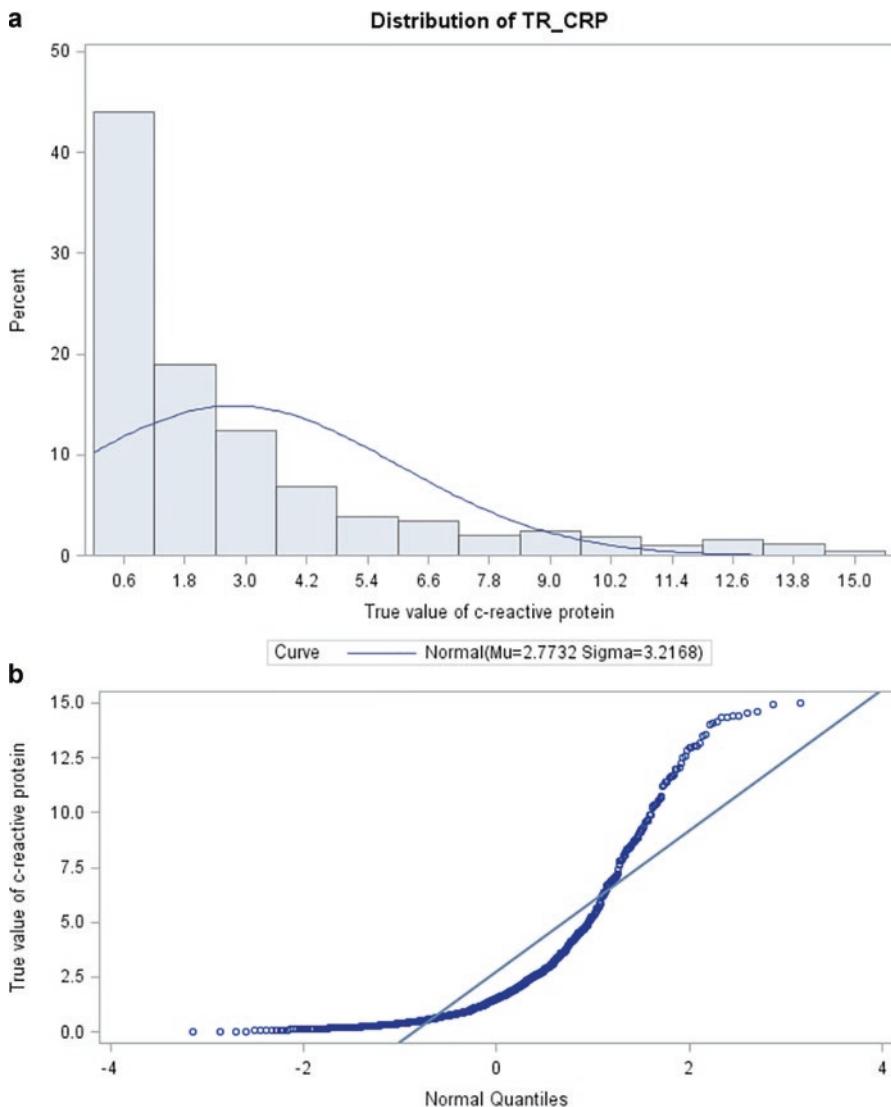


Fig. 22.1 Quantile plot and histogram of the dependent variable

### 1.3.2 Regression Diagnostics [7–9]

#### Influence Statistics

Cook's D is designed to measure the effect on coefficients and provide the influence of a particular data point. A value close to 0 indicates no influence; the larger the value, the larger the influence. The student residual contains the z-scores for the residual. A residual larger than +3 or less than -3 indicates an outlier. Similarly,

**Table 22.2** Partial display of influential statistics

Obs	Dependent Variable	Predicted Value	Output Statistics														DFBETAS						
			Std Error Predict			Student Residual			Cook's D	RStudent	Hat Diag H		Cov Ratio		DFFITS	Intercept	bwkgt	RACE	SEX	age	BMI	cursmk	
			Residual	Std Error Residual	Student Residual																		
1	0.30	2.1528	0.3019	-1.8528	2.937	-0.631	0.001	-0.6306	0.0105	1.0161	-0.0648	-0.0129	-0.0241	0.0078	-0.0217	0.0392	0.0169	-0.0238					
2	0.22	1.3152	0.2123	-1.0952	2.945	-0.372	0.000	-0.3717	0.0052	1.0131	-0.0268	-0.0025	-0.0068	0.0040	0.0138	-0.0036	0.0071	0.0135					
3	1.37	2.4298	0.2313	-1.0598	2.944	-0.360	0.000	-0.3598	0.0061	1.0142	-0.0283	0.0068	-0.0017	0.0064	-0.0128	-0.0114	0.0079	-0.0127					
4	9.61	0.8501	0.3410	8.7599	2.933	2.987	0.017	3.0024	0.0133	0.9423	0.3491	0.0383	-0.0661	0.1939	-0.1509	0.0764	-0.1666	0.0978					
5	1.49	1.8565	0.3553	-0.3665	2.931	-0.125	0.000	-0.1250	0.0145	1.0239	-0.0151	0.0059	-0.0020	-0.0006	-0.0037	-0.0074	0.0074	-0.0038					
6	1.57	4.2147	0.2482	-2.6447	2.942	-0.899	0.001	-0.8988	0.0071	1.0089	-0.0758	0.0340	-0.0011	0.0255	-0.0330	-0.0414	-0.0307	0.0247					
7	3.59	3.3401	0.2658	0.2499	2.941	0.085	0.000	0.0849	0.0081	1.0174	0.0077	-0.0016	-0.0023	-0.0026	0.0028	0.0050	0.0003	-0.0028					
8	2.38	1.2875	0.2286	1.0924	2.944	0.371	0.000	0.3709	0.0060	1.0140	0.0286	0.0029	0.0131	-0.0027	-0.0134	-0.0035	-0.0068	-0.0133					
9	7.58	1.9259	0.3429	5.6541	2.933	1.928	0.007	1.9314	0.0135	0.9887	0.2258	-0.1378	0.1553	0.0062	0.0897	0.0862	-0.0632	0.0601					
10	12.01	2.7690	0.2242	9.2410	2.944	3.139	0.008	3.1571	0.0058	0.9271	0.2405	0.0566	-0.1077	-0.0993	0.0910	0.0544	-0.0476	-0.1060					
11	0.98	2.1190	0.2234	-1.1390	2.944	-0.387	0.000	-0.3866	0.0057	1.0136	-0.0293	0.0057	-0.0074	0.0052	0.0136	-0.0117	-0.0026	0.0127					
12	1.12	2.5849	0.2762	-1.4649	2.940	-0.498	0.000	-0.4981	0.0087	1.0158	-0.0468	0.0142	-0.0045	-0.0301	-0.0133	-0.0170	0.0185	0.0182					
13	0.41	2.0719	0.3094	-1.6619	2.936	-0.566	0.001	-0.5657	0.0110	1.0174	-0.0596	-0.0111	0.0008	-0.0341	-0.0125	0.0204	0.0294	-0.0204					
14	1.53	3.7380	0.2100	-2.2080	2.945	-0.750	0.000	-0.7495	0.0051	1.0091	-0.0534	-0.0070	-0.0034	0.0212	-0.0243	0.0190	-0.0150	0.0202					
15	1.63	3.2209	0.2646	-1.5909	2.941	-0.541	0.000	-0.5407	0.0080	1.0146	-0.0486	-0.0036	-0.0131	0.0106	0.0200	0.0148	-0.0237	0.0132					
16	12.58	5.3170	0.2894	7.2630	2.938	2.472	0.008	2.4800	0.0096	0.9534	0.2442	-0.1268	0.0506	0.1234	0.0734	0.0353	0.1311	-0.0518					

larger values of standardized residual and DFFITS (which provide the measure on  $\hat{y}$ ) and COV ratio indicate an outlier. An observation is influential if its deletion leads to major changes in the fitted regression. Cook's D is designed to measure the effect on coefficients and provides the influence of a particular data point. A value close to 0 indicates no influence; the larger the value, the larger the influence. DFBETAS measures the influence of the  $i$ th (an individual) observation on a single coefficient of beta (Table 22.2).

### Multicollinearity

The diagnosis of multicollinearity depends on the following criteria:

1. Independent variables are highly related ( $R > 0.85$ ).
2. Substantial  $R$  square and statistically insignificant coefficients.
3. Beta coefficients and standard error are large.
4. Signs are unexpected.

### 1.3.3 Collinearity Diagnosis and SAS Output

The variance inflation factor (VIF) is calculated from the correlation matrix of the independent variables. If VIF is  $>10$ , it is considered severe collinearity. There are two types of collinearity: nonessential collinearity and essential collinearity. The first collinearity is with the intercept, and the second collinearity is with other independent variables. The *condition number* of a matrix is defined as the ratio of the largest singular value (eigenvalue) to the smallest singular value. The large condition number ( $>30$ ) indicates that the matrix is poorly conditioned and the variable has a dependency (Table 22.3). The *condition index*, an extension of the VIF, provides information on each dimension of the matrix. The *condition indices* of 30–100

**Table 22.3** Collinearity diagnosis statistics

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t	Variance Inflation
Intercept	Intercept	1	-3.55551	1.43173	-2.48	0.0132	0
bwkg1	Birth weight in kg	1	-0.39655	0.23126	-1.71	0.0868	1.10729
RACE	Black vs. White	1	0.11354	0.24980	0.45	0.6496	1.11404
SEX	Female vs Male	1	1.22554	0.21847	5.61	<.0001	1.03454
age	Age in years	1	0.01419	0.02408	0.59	0.5559	1.01124
BMI	Body mass index	1	0.17050	0.01615	10.56	<.0001	1.06250
cursmk	Currently smoking	1	0.35079	0.21802	1.61	0.1080	1.03342

Collinearity Diagnostics (intercept adjusted)							
Number	Eigenvalue	Condition Index	Proportion of Variation				
			bwkg1	RACE	SEX	age	BMI
1	1.40762	1.00000	0.18560	0.23325	0.10782	0.02501	0.05646
2	1.17977	1.09230	0.12328	0.00369	0.06638	0.04891	0.32296
3	1.00449	1.18378	0.03591	0.04016	0.16112	0.68943	0.02983
4	0.91038	1.24346	0.00293	0.19484	0.25349	0.09250	0.05657
5	0.84019	1.29436	0.15745	0.00002538	0.40266	0.13394	0.26264
6	0.65755	1.46312	0.49481	0.52803	0.00853	0.01021	0.27154
							0.07451

indicate moderate to severe dependencies, and a value larger than 100 indicates serious collinearity. The proportion of variation provides further information on collinearity if the proportions are greater than 50% between two variables.

### SAS Code for Multiple Linear Regression

```
proc reg;
model crpl=bwkg1 race sex age bmi cursmk;
run;
```

#### 1.3.4 Model Fitness Evaluation

##### Global F-Test

First, we assess the overall model with the *F* test; if the *F*-value is large and the *p*-value is <0.05, we can say there is a significant relationship between the dependent and independent variables. The results are obtained from the analysis of the

**Table 22.4** ANOVA statistics

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
<b>Model</b>	6	242.26541	40.37757	33.34	<.0001
<b>Error</b>	766	927.61636	1.21099		
<b>Corrected Total</b>	772	1169.88177			

variance table (Table 22.4). In multiple linear regression, *F*-values provide the overall fitness of the model and indicate that all the independent variables, taken together, contribute to the dependent variable.

### Hypothesis for Model Fitness

$H_0$ : Model does not fit well with independent variables or  $H_0: B1 = B2 = B3 = 0$ .  
 $H_a$ : Model provides well with independent variables, or at least one beta is not zero.

If the *p*-value <0.05 from the *F* test, we can say that model fits well with the independent variables in the model (Table 22.4).

### Effect of the Individual Variable from the Parameter Estimates

Although the *F*-test provides model fitness, to check the independent effect, we need to check the *t*-test using the parameter estimates. The hypotheses are as follows:

$H_0$ : There is no relationship between birth weight and hs-CRP in asymptomatic young adults adjusting for race, sex, age, bmi, and cursmk.  
 $H_a$ : There is a significant relationship between birth weight and hs-CRP in asymptomatic young adults adjusting for race, sex, age, bmi, and cursmk.

### Interpretation of coefficient of determination (*R*<sup>2</sup>)

Based on Table 22.5,  $R^2 = 20.71$ , which means 20.71% of the variability of hs-CRP was explained by birth weight, sex, and BMI. Now, let us examine Table 22.6 data – birthweight, sex, BMI, and cursmk (current smoking) are significant predictors of hs-CRP.

### Conclusion

There is a relationship between birth weight and hs-CRP in asymptomatic younger adults after adjusting for race, sex, age, BMI, and current smoking.

### Interpretation of Parameter Estimates

#### For Continuous Variable

- With one unit increase in birth weight, the mean level of log CRP will decrease by 0.233 units, *p*-value = 0.007.

**Table 22.5** Displayed the  $R^2$ 

<b>Root MSE</b>	1.10045	<b>R-Square</b>	0.2071
<b>Dependent Mean</b>	0.36580	<b>Adj R-Sq</b>	0.2009
<b>Coeff Var</b>	300.83504		

**Table 22.6** Displayed the parameter estimates of hs-CRP

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
<b>Intercept</b>	Intercept	1	-2.26591	0.53360	-4.25	<.0001
<b>bwkg1</b>		1	-0.23345	0.08619	-2.71	0.0069
<b>RACE</b>	Race	1	-0.01193	0.09310	-0.13	0.8980
<b>SEX</b>	Sex	1	0.43269	0.08142	5.31	<.0001
<b>age</b>		1	0.01304	0.00898	1.45	0.1467
<b>BMI</b>		1	0.07608	0.00602	12.64	<.0001
<b>cursmk</b>		1	0.18631	0.08125	2.29	0.0221

- With one unit increase in BMI, the mean level of log CRP will increase by 0.076 units, with a  $p$ -value <0.0001.
- Females had higher values of the average 0.433 unit log CRP compared to males (females were coded 2 and males coded 1),  $p$ -value <0.0001.
- Smokers had higher values (0.1863) of hs-CRP compared to nonsmokers,  $p$ -value = 0.02.

## 2 Model Selection Criteria for Regression

You should run a model selection procedure to identify the important variables and remove redundant variables. Three types of model selections are as follows: forward method, backward method, and stepwise method. In the forward selection method, variables are sequentially entered into the model. In the backward selection method, all variables are put into the model, and steps are taken to select the significant variables. In the stepwise selection method, independent variables are used stepwise, one at a time; the significant variables are kept in the model, and another variable is used, thus selecting all the variables in the model. The stepwise selection method is a combination of sequential and backward elimination. Therefore, it is sometimes considered the best model selection.

There are certain criteria for choosing the best model [6].

- **Coefficient of determination ( $R^2$ ):** It ranges from 0 to 1. An  $R^2$  of 1 indicates that the regression predictions perfectly fit the data.
- **Adjusted  $R^2$ :** It is a corrected goodness-of-fit (model accuracy) measure for linear models. An adjusted  $R^2$  considers and tests different independent variables against the model, whereas  $R^2$  does not.  
As  $R^2$  increases with adding variables, adjusted  $R^2$  reduces the value after giving a penalty for adding additional variables.
- **Standard error of the regression mean:** The value of the SEM should be low.
- **Covariance structure of residual:** The likelihood-based information criteria, known as the Akaike Information Criterion (AIC), are used. AIC estimates the quality of each model relative to each of the other model.  
 $AIC = -2 \ln(L) + 2k$ , where  $L$  = likelihood and  $k$  = no. of parameters.  
AIC is low for models with high log-likelihoods. This means the model fits the data better.
- **Bayesian Information Criterion (BIC):** BIC is a method of scoring and selecting a model.  
 $BIC = -2 * LL + \log(N)k$ , where  $LL$  = log-likelihood of the model,  $N$  = sample size, and  $k$  = number of parameters.  
Lower values of BIC (like AIC) mean a good fit with the data set.
- **Mallows's  $C_p$  statistic:** It estimates the magnitude of bias in the dependent variable. If  $C_p$  is close to the number of parameters plus the intercept, we can say that the model is good with the lowest bias. Larger values of  $C_p$  indicate that the important variable has been left out.  $C_p$  is often used as a stopping rule for various forms of stepwise regression.

## 2.1 SAS Code for Model Selection Procedure [10]

```

proc reg;
  model crpl=bwkg1 race sex age bmi cursmk / selection=forward sle=.15;
  run;
proc reg;
  model crpl=bwkg1 race sex age bmi cursmk / selection=backward sls=.15;
  run;
proc reg;
  model crpl=bwkg1 race sex age bmi cursmk / selection=stepwise sle=.15 sls=.15;
  run;
*proc glmselect;
model crpl=bwkg1 race sex age bmi cursmk / selection=forwarde stats= bic details=all;
run;

```

\*Note: proc glmselect option was used to get aic bic as proc reg provided only Cp

## 2.2 Partial SAS Output from Model Selection Procedure (Tables 22.7a, 22.7b, and 22.7c)

In this model selection process, we can say that all models are good as they provide similar information (Table 22.8).

## 3 Interaction or Effect Modification

Interaction occurs when independent variables interact and the overall effect differs from their effect [11, 12]. If the effect is greater than its combined effect, it is called a synergistic interaction. If the effect is smaller than their combined effect, it is called antagonistic interaction. Interactions are commonly represented in multiple regression models, including a term consisting of the product of two of the independent variables. For example, we could add the interaction term (race \* sex) to the model.

$$Y (\text{crpl}) = b_0 + b_1 \text{birthweight1} + b_2 \text{race} + b_3 \text{sex} \\ + b_4 \text{race}^* \text{sex} + b_5 \text{age} + b_6 \text{bmi} + b_7 \text{cursmk}$$

Here, you are testing for the slope of interaction variable is race \* sex.

$H_0$ :  $b4 = 0$ , given birth weight, race, sex, age, bmi, and cursmk are in the model  
 $H_a$ :  $b4 \neq 0$  given birth weight, race, sex, age, bmi, and cursmk are in the model.

If the coefficient associated with the interaction term is significantly different from zero, it indicates that there is a significant interaction between the independent variables; the effect of one depends on the value of the other values.

### SAS code for interaction

```
proc reg;
model crpl=bwkg1 race sex age bmi cursmk race*sex;
run;
```

## 4 Binary Logistic Regression

Binary logistic regression is applied when the dependent variable is the dichotomous response, for example, disease: yes versus no; dead versus alive; any condition (yes vs. no), low birth weight versus normal birth weight. The coding of variables is critical. In our dataset, as we demonstrated multiple logistic regression for predicting low birth weight, the dependent variable is the birth weight (low vs. normal birth weight). Independent variables are used in the model as either

**Table 22.7a** Forward procedure

Effects: Intercept bwkg1 SEX BMI				
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	233.00347	77.66782	63.75
Error	769	936.87830	1.21831	
Corrected Total	772	1169.88177		

Root MSE	1.10377
Dependent Mean	0.36580
R-Square	0.1992
Adj R-Sq	0.1960
AIC	931.62807
AICC	931.70630
BIC	158.62137
C(p)	8.64826
SBC	175.22919

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-1.655099	0.364081	-4.55
bwkg1	1	-0.225007	0.083246	-2.70
SEX	1	0.421440	0.081295	5.18
BMI	1	0.073126	0.005863	12.47

**Table 22.7b** Backward procedure

Effects: Intercept bwkg1 SEX BMI				
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	233.00347	77.66782	63.75
Error	769	936.87830	1.21831	
Corrected Total	772	1169.88177		

Root MSE	1.10377
Dependent Mean	0.36580
R-Square	0.1992
Adj R-Sq	0.1960
AIC	931.62807
AICC	931.70630
BIC	158.62137
C(p)	8.64826
SBC	175.22919

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-1.655099	0.364081	-4.55
bwkg1	1	-0.225007	0.083246	-2.70
SEX	1	0.421440	0.081295	5.18
BMI	1	0.073126	0.005863	12.47

**Table 22.7c** Stepwise procedure

Effects: Intercept bwkg1 SEX BMI				
Analysis of Variance				
Source	DF	Sum of Squares	Mean Square	F Value
Model	3	233.00347	77.66782	63.75
Error	769	936.87830	1.21831	
Corrected Total	772	1169.88177		

Root MSE	1.10377
Dependent Mean	0.36580
R-Square	0.1992
Adj R-Sq	0.1960
AIC	931.62807
AICC	931.70630
BIC	158.62137
C(p)	8.64826
SBC	175.22919

Parameter Estimates				
Parameter	DF	Estimate	Standard Error	t Value
Intercept	1	-1.655099	0.364081	-4.55
bwkg1	1	-0.225007	0.083246	-2.70
SEX	1	0.421440	0.081295	5.18
BMI	1	0.073126	0.005863	12.47

**Table 22.8** Summary of findings of the model selection

Procedure	Variables remained in the model	Adjusted R-square	Standard error of the estimate= (MSE)	$C_p$	AIC	BIC
Forward	Bwkg1, sex, and bmi	0.1960	1.2183	8.64	931.63	158.62
Backward	Bwkg1, sex, and bmi	0.1960	1.2183	8.64	931.63	158.62
Stepwise	bwkg1, sex, and bmi	0.1960	1.2183	8.64	931.63	158.62

AIC Akaike Information Criterion, BIC Bayesian Information Criterion

categorical, ordinal, or continuous variables. Therefore, we are modeling race, sex, and current smoking as categorical variables and BMI and hs-CRP as continuous variables. Previously, we demonstrated that we used the log value of hs-CRP (crpl) in the linear regression model as hs-CRP is not normally distributed. Our dependent variable is qualitative/binary, so we could not assume a linear function like a linear regression model. Therefore, we cannot apply the least squares method to estimate the parameters. Instead, we can estimate the coefficients using the maximum likelihood function (MLE). In MLE, coefficients are chosen to maximize the probability of  $y$  given  $x$  (likelihood). The computer uses an “iterations” process such as Fisher scoring to get the maximum likelihood estimates. Fisher scoring is one standard method for the iterative method in MLE [13, 14]. The criteria are below for model fit assumptions for applied Logistic Regression [15].

1. The logit transformation of the dependent/outcome variable has to have a linear relationship with the independent or predictor variables.
2. There is no multicollinearity problem.
3. There should be no influential problem.
4. There should be a large sample size, at least 10 events per independent variable.

## 5 Model Equation for Multiple Logistic Regression

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = b_0 + b_1 \text{ race} + b_2 \text{ sex} + b_3 \text{ age} + b_4 \text{ bmi} + b_5 \text{ cursmk} + b_6 \text{ crpl}$$

where  $p$  = probability of low birth weight,  $\beta_0$  = intercept,  $\beta_1$  = slope for the race, and so on.

We could test whether the independent variable is associated with low birth weight as  $H_0: \beta_1 = 0$ ;  $H_a: \beta_1 \neq 0$  for race controlling for all other variables in the model and so on.

For interaction, we created an interaction term in the model, which is the product of two variables. Then we can test the coefficient of interaction [16].

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = b_{0+} b_1 \text{ race} + b_2 \text{ sex} + b_3 \\ age + b_4 \text{ bmi} + b_5 \text{ cursmk} + b_6 \text{ crpl} + b_7 \text{ cursmk*crpl}$$

$H_0: b_7 = 0$ ;  $H_a: b_7 \neq 0$  for the interaction of “cursmk” and “crpl” controlling for all other variables.

As in Table 22.9, the  $\beta_6$  of the interaction of crp\*cursmk was not significant. We removed the interaction term from the model.

## 5.1 SAS Code for Multiple Logistic Regression with Interaction

```
proc logistic descending;
class race sex cursmk/param=ref;
label
lowbirth= " low birth vs. normal birth"
race= "1 for White and 2 for Black"
sex = "1 for male and 2 for female"
age = "Age in years"
bmi = "Body mass index"
cursmk= "Currently smoking status (1= yes vs. 0=no)"
crpl=" log value of hs-CRP";
model lowbirth=race sex age bmi cursmk crpl cursmk*crpl/influence corrb;
run;
```

## 5.2 SAS Code for Binary Logistic Regression

In the SAS program, the PROC LOGISTIC procedure is used to analyze data when the dependent variable is categorical, containing two levels. Birth weight was categorized into normal and low birth weight, coded as low birth weight as 1 and normal birth weight as 0. SAS code for model assumption in the model as option influence for influential statistic and corrb for collinearity diagnosis. Of note, there was no option for collinearity checking in logistic regression. The corrb option provides a correlation matrix to check collinearity. The correlation matrix is provided below in Table 22.10.

The correlation matrix did not provide information on the high correlation between independent variables. Regression diagnostics in the data set did not show high leverage in Table 22.11. It is the measure between the  $x$  values for the  $i$ th case and the means of the  $x$  values for all  $n$  cases. Regarding influential statistics, according to Belsley, D.A. Hat matrix diagonal  $> \frac{2p'}{n}$  ( $n$  = case and  $p'$  = number of

**Table 22.9** Displayed the interaction term

Joint Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
RACE	1	6.3493	0.0117
SEX	1	6.1971	0.0128
age	1	0.0114	0.9151
BMI	1	5.2330	0.0222
cursmk	1	0.1214	0.7275
crpl	1	4.0112	0.0452
crpl*cursmk	1	0.4743	0.4910

parameters) indicates high leverage points. DFBETAS identifies the influential observations for the individual regression coefficients. Belsley, D.A., suggested the cutoff point of  $2/\sqrt{n}$ . Values of DFBETAS greater than two are considered a major point of influence [3, 4]. The Pearson residual is the difference between observed and fitted values divided by an estimate of the standard deviation of the observed value displayed in Fig. 22.2. The deviance residual is based on the deviance or likelihood ratio chi-squared statistic. Booth person residual and deviance residual exceeding two may indicate a lack of fit in the model [15, 17, 18].

### 5.3 SAS Code for Logistic Analysis for Multiple Logistic Regression Without Interaction [19]

Here, the interaction term was removed as it was not significant.

```
proc logistic descending;
class race (ref='1') sex (ref='1') cursmk (ref='0');
model lowbirth=race sex age bmi cursmk crpl;
run;
quit;
```

**Table 22.10** Displayed correlation matrix

Estimated Correlation Matrix							
Parameter	Intercept	RACE1	SEX1	age	BMI	cursmk0	crpl
Intercept	1.0000	-0.1151	0.0255	-0.8408	-0.4190	-0.1212	0.0725
RACE1	-0.1151	1.0000	-0.0192	-0.1014	0.1741	0.0049	-0.0088
SEX1	0.0255	-0.0192	1.0000	-0.0600	-0.0492	0.0345	0.1151
age	-0.8408	-0.1014	-0.0600	1.0000	-0.0793	0.0910	0.0125
BMI	-0.4190	0.1741	-0.0492	-0.0793	1.0000	-0.2104	-0.3297
cursmk0	-0.1212	0.0049	0.0345	0.0910	-0.2104	1.0000	0.0324
crpl	0.0725	-0.0088	0.1151	0.0125	-0.3297	0.0324	1.0000

## 6 Model Fit Evaluation/Model Fit Statistics

- AIC, SC, and  $-2 \log L$  values:** With lower values of AIC, SC, and  $-2 \log L$  in the model with predictors than those with the intercept-only model, we could say that the model with covariates fits well [18]. The AIC, SC, and  $-2 \log L$  assess the hypothesis that the model with predictor variables fits better than the model without predictor variables or the null model. Model fit statistics are provided from the SAS output in Table 22.12.
- In the likelihood ratio chi-square test** 26.3824 and  $p$ -value = 0.002, we rejected the null hypothesis and concluded that adding variables improved the model in Table 22.13. The improvement test is like the global  $F$ -test in the linear regression model. The score and Wald tests provide the same information.  
Hypothesis (goodness of fitness)

$H_0$ : The simple/null model explains data better

$H_a$ : Complex model explains data better

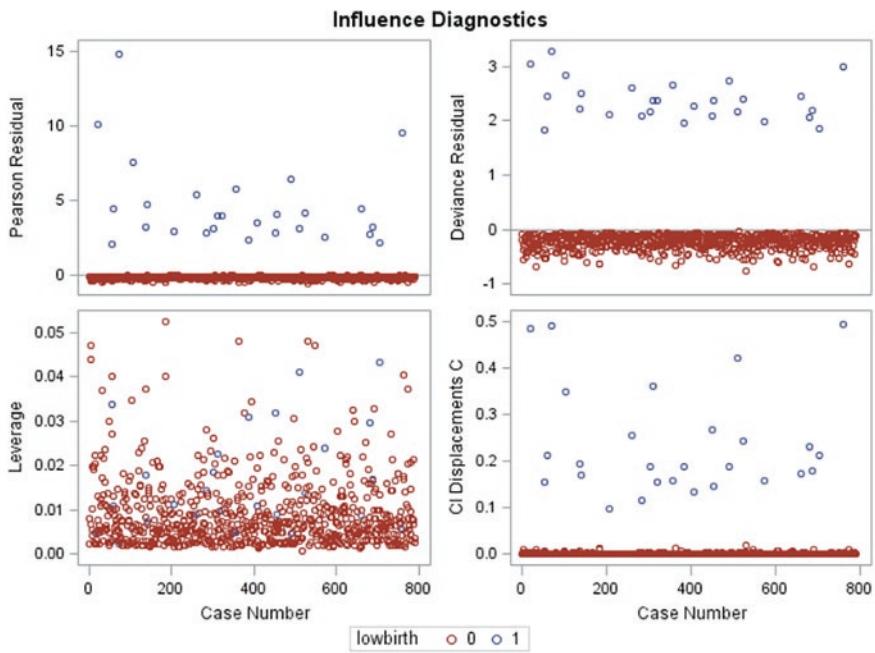
### 3. Model Evaluation with ROC Curve

## 7 SAS Code for ROC

```
proc logistic descending;
class race (ref='1') sex (ref='1') cursmk (ref='0');
model lowbirth=race sex age bmi cursmk crpl/rsq outroc=roc;
run;
quit;
```

**Table 22.11** Displayed regression diagnostic

Case Number	Regression Diagnostics																
	Covariates																
	Black vs. White 1	Female vs Male 1	Age in years	Body mass index	Currently smoking 0	log value of hs-CRP	Pearson Residual	Deviance Residual	Hat Matrix Diagonal	Intercept DFBeta	RACE1 DFBeta	SEX1 DFBeta	age DFBeta	BMI DFBeta	cursmk0 DFBeta	crpl DFBeta	
1	1.0000	0	28.6000	23.0376	0	-1.2040	-0.1456	-0.2049	0.00795	-0.00705	-0.00450	0.00215	0.00742	-0.00185	0.00578	0.00713	
2	1.0000	1.0000	37.6000	26.2089	1.0000	-1.5141	-0.0671	-0.0948	0.00236	0.000378	-0.00083	-0.00225	-0.00016	-0.00029	-0.00082	0.00154	
3	1.0000	0	40.2000	22.6122	0	0.3148	-0.2005	-0.2808	0.00684	0.00268	-0.00614	0.00393	-0.00639	0.00228	0.00847	0.00203	
4	0	1.0000	39.9000	18.6778	0	2.2628	-0.2621	-0.3645	0.0439	0.00593	0.0194	-0.0391	-0.0114	0.0278	0.0107	-0.0285	
5	0	0	43.3000	17.9104	0	0.3988	-0.4189	-0.5686	0.0471	0.0106	0.0461	0.0143	-0.0558	0.0484	0.0270	-0.0349	
6	1.0000	0	41.3000	35.0132	1.0000	0.4511	-0.1508	-0.2120	0.00540	0.00760	-0.00513	0.00240	-0.00487	-0.00604	-0.00274	0.00346	
7	1.0000	0	43.1000	28.8113	1.0000	1.2782	-0.2254	-0.3148	0.0101	0.0149	-0.00892	0.00412	-0.0155	0.00042	-0.00879	-0.00332	
8	1.0000	1.0000	35.2000	26.7741	1.0000	0.8671	-0.1021	-0.1440	0.00429	-0.00030	-0.00201	-0.00577	0.000688	0.000516	-0.00189	-0.00087	
9	1.0000	0	41.5000	21.7571	0	2.0255	-0.2664	-0.3970	0.0198	0.00636	-0.0111	0.00518	-0.0176	0.0151	0.0151	-0.0209	
10	1.0000	0	38.2000	25.6383	1.0000	2.4857	-0.3177	-0.4386	0.0193	0.00573	-0.0172	0.00378	-0.0104	0.0156	-0.0178	-0.0301	
11	1.0000	1.0000	40.0000	30.7890	1.0000	-0.0202	-0.0744	-0.1050	0.00247	0.00122	-0.00109	-0.00286	-0.00073	-0.00070	-0.00092	0.00069	

**Fig. 22.2** Influence diagnostics

The receiver operative curve (ROC) provided model evaluation with a plot of sensitivity over 1-specificity. Figure 22.3 displayed the area under the curve as 0.7757, which is good as it is close to 1, indicating good fit, and 0.50, indicating poor fit [19].

**Table 22.12** Model fit statistics

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	236.186	221.803
SC	240.836	254.355
-2 Log L	234.186	207.803

**Table 22.13** Improvement test

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	26.3824	6	0.0002
Score	25.1047	6	0.0003
Wald	20.9478	6	0.0019

## 7.1 Model Evaluation for Continuous Variable: Hosmer Lemeshow Test

When independent variables are continuous, SAS code in the model option lackfit provides Hosmer and Lemeshow results for the goodness of fitness statistics ( $L^2 = 2e^{\text{observe}} \ln(\text{observe}/\text{Expected})$ ). SAS computes a chi-square from the observed and expected values. Large chi-square values and a small  $p$ -value indicate a lack of fit in the model [20].

As Hosmer and Lemeshow's goodness of fitness, chi-square has a small value of 6.2844 and a large  $p$ -value of 0.6154, we fail to reject the null hypothesis. That means the model has a good predictive value in Table 22.14.

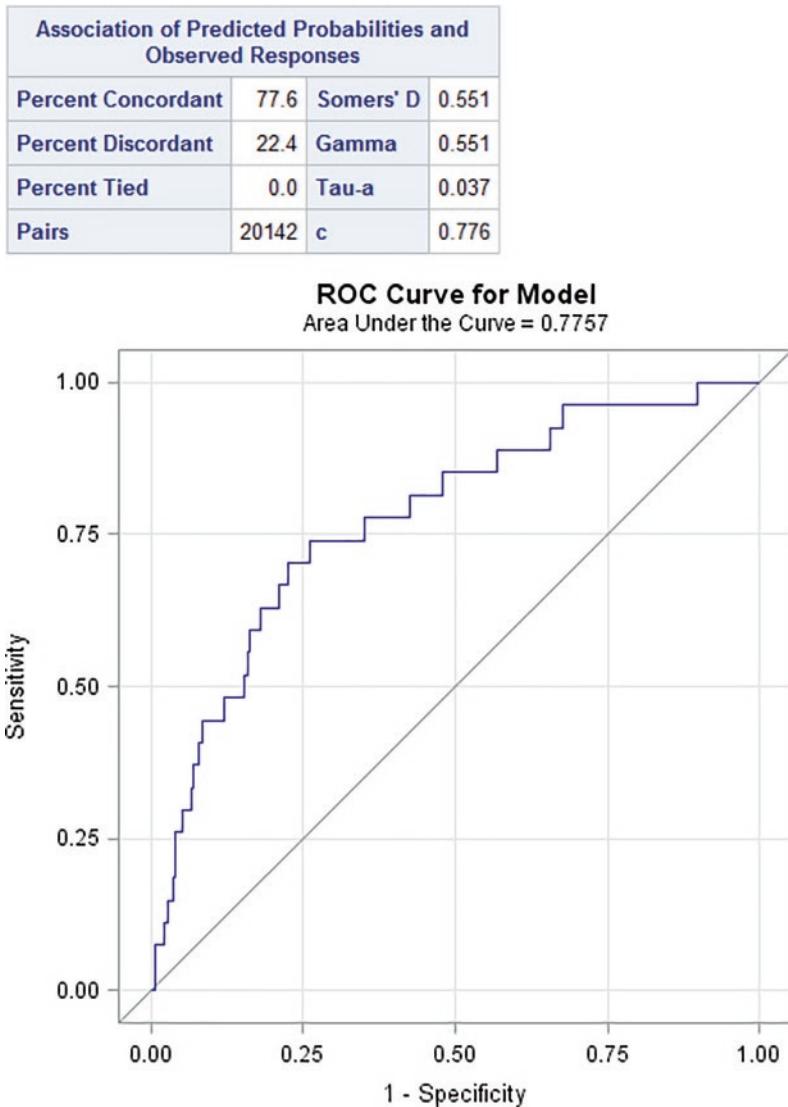


Fig. 22.3 Displayed ROC for model

## 7.2 SAS Code for Hosmer and Lemeshow Test

```
proc logistic descending data=final;
model lowbirth=age bmi crpl/lackfit;
run;
```

**Table 22.14** Displayed Hosmer–Lemeshow statistics

Partition for the Hosmer and Lemeshow Test						
Group	Total	lowbirth = 1		lowbirth = 0		
		Observed	Expected	Observed	Expected	
1	78	1	0.92	77	77.08	
2	78	0	1.30	78	76.70	
3	78	2	1.59	76	76.41	
4	78	3	1.85	75	76.15	
5	78	4	2.13	74	75.87	
6	78	0	2.44	78	75.56	
7	78	3	2.82	75	75.18	
8	78	4	3.34	74	74.66	
9	78	5	4.27	73	73.73	
10	74	5	6.35	69	67.65	

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
6.9853	8	0.5382

## 8 R Square Statistics

Max-rescaled  $R$ -square 0.1284 indicates that the model contributes to a 12.84% prediction of low birth weight by those variables in Table 22.15.

## 9 Interpretation of Regression Parameter

Like the  $t$ -test in linear regression, the **Wald Chi-square test** provides the independent or adjusted effect. Wald test = square of  $z$  test =  $(\text{beta}/\text{standard error of beta})^2$  in Table 22.16.

**Table 22.15** Displayed R-square statistics

R-Square	0.0336	Max-rescaled R-Square	0.1284
----------	--------	-----------------------	--------

**Table 22.16** Maximum likelihood estimates

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.7553	1.8114	0.9390	0.3325
RACE	2	1	0.5306	0.2046	6.7219	0.0095
SEX	2	1	0.7825	0.3129	6.2564	0.0124
age		1	0.00323	0.0443	0.0053	0.9419
BMI		1	-0.0788	0.0347	5.1513	0.0232
cursmk	1	1	-0.1759	0.2127	0.6842	0.4082
crpl		1	0.3747	0.1825	4.2163	0.0400

### *Interpretation of estimate with p-value*

#### (a) For categorical variables

Compared to Whites, Blacks were associated with higher log odds of low birth weight by 0.5306. It is significant as the *p*-value = 0.0095.

Compared to males, females had higher log odds of low birth weight by 0.7825, which is significant as the *p*-value = 0.0124.

#### (b) For the continuous variables

For every unit increase in age, the log odds of having low birth weight decrease by 0.0032, but it is non-significant as the *p*-value = 0.9419.

For every unit increase in BMI, the log odds of having low birth weight decrease by -0.0788, and it is significant as the *p*-value = 0.023.

For every unit increase in the log value of CRP, the log odds of having low birth weight decrease by 0.3747, and it is significant as the *p*-value = 0.040.

**Table 22.17** Odds Ratio statistics

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
RACE 2 vs 1	2.890	1.296	6.445
SEX 2 vs 1	4.783	1.403	16.304
age	1.003	0.920	1.094
BMI	0.924	0.863	0.989
cursmk 1 vs 0	0.703	0.306	1.619
crpl	1.455	1.017	2.080

### Interpretation of ORs with and 95% CI from Table 22.17

As the odds ratio is the exponential coefficient, it is interpreted as

#### (a) For categorical variables

Compared to Whites, Blacks were associated with higher odds of low birth weight by 2.890, which is significant as the 95% CI does not include 1.

Compared to males, females had higher odds of low birth weight by 4.783, which is significant.

#### (b) For continuous variables

For every unit increase in age, the odds of having low birth weight decrease by 1.003, but it is nonsignificant as the 95% CI includes 1.

For every unit increase in BMI, the odds of having low birth weight decrease by 0.924, and it is significant as the 95% CI does not include 1.

For every unit increase in the log value of hs-CRP, the odds of having low birth weight decrease by 1.455, and it is significant as the 95% CI does not include 1.

#### (c) For Ordinal Predictor and Trend Analysis:

A SAS code was provided if you want to use an ordinal variable and dose-response relationship. For example, the BMI class has three levels: normal weight, overweight, and obese. First of all, BMI class was treated as a continuous variable, showing that the quadratic trend was significant, as shown in Table 22.18.

**Table 22.18** Displayed trend analysis statistics

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.1996	1.5239	11.6415	0.0006
RACE	1	1	-1.0473	0.4001	6.8514	0.0089
SEX	1	1	-1.9071	0.6228	9.3784	0.0022
bmiclass		1	3.6604	1.6250	5.0744	0.0243
bmiclass*bmiclass		1	-0.9352	0.3980	5.5222	0.0188

Then, BMI class was treated as an ordinal predictor. The results also showed the same consistent finding with orthogonal (polynomial) coding. Class option orthpoly will provide linear quadratic and cubic trends for ordinal predictors to examine the dose-response relationship in SAS. Here, the BMI class has three levels: normal, overweight, and obese. SAS created two orthogonal (trend) coefficients for linear and quadratic components [21]

## 10 SAS Code for Interaction

```
proc logistic descending;
class race sex/param=ref;
model lowbirth=race sex bmiclass bmiclass*bmiclass;
run;
```

## 11 SAS Code for Trend Analysis

```
proc logistic descending data=final;
class race sex bmiclass/param =orthpoly;
model lowbirth=race sex bmiclass;
run;
```

The SAS output showed that the linear trend was insignificant ( $p = 0.7652$ ), but the quadratic trend was significant as the  $p$ -value = 0.0188, as shown in Table 22.19.

**Table 22.19** Trend analysis for ordinal predictor

Class Level Information			
Class	Value	Design Variables	
RACE	1	-1.000	
	2	1.000	
SEX	1	-1.000	
	2	1.000	
bmiclass	Normal weight	-1.225	0.707
	Obese	1.225	0.707
	Overweight	0	-1.414

Model Convergence Status			
Convergence criterion (GCONV=1E-8) satisfied.			

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-3.7202	0.3176	137.1949	<.0001
RACE	OPOLY1	1	0.5237	0.2001	6.8514	0.0089
SEX	OPOLY1	1	0.9536	0.3114	9.3784	0.0022
bmiclass	OPOLY1	1	-0.0656	0.2197	0.0892	0.7652
bmiclass	OPOLY2	1	-0.4409	0.1876	5.5222	0.0188

## 12 SAS Code for Forward, Backward, and Stepwise Logistic Regression

```
proc logistic descending;
class race sex cursmk/param=ref;
model lowbirth=race sex age bmi cursmk crpl/selection=forward;
run;
proc logistic descending;
class race sex cursmk/param=ref;
model lowbirth=race sex age bmi cursmk crpl/selection=backward;
run;

proc logistic descending;
class race sex cursmk/param=ref;
model lowbirth=race sex age bmi cursmk crpl/selection=stepwise;
run;
```

## 13 Further Practice

Multiple choice questions:

1. Multiple linear or logistic regression is widely used because
  - (a) It allows to quantify effect
  - (b) to examine the correlation
  - (c) to show independent effect
  - (d) All of the above
2. The dependent variable in linear regression is
  - (a) Continuous variable
  - (b) Categorical variable
  - (c) Ordinal variable
  - (d) All
3. The independent variable in linear regression is
  - (a) Continuous variable
  - (b) Categorical variable
  - (c) Ordinal variable
  - (d) All of the above
4. The dependent variable in logistic regression is
  - (a) Continuous variable
  - (b) Categorical variable
  - (c) Ordinal variable
  - (d) All of the above

5. The independent variable in logistic regression is
  - (a) Continuous variable
  - (b) Categorical variable
  - (c) Ordinal variable
  - (d) All of the above
6. The data fit assumption is evaluated
  - (a) Normality of variable
  - (b) Influential statistics
  - (c) Collinearity check
  - (d) All of the above
7. The model fitness is evaluated by either linear or logistic regression
  - (a) Global  $F$ -test
  - (b) Likelihood ratio test
  - (c) Hosmer Lemeshow test
  - (d) All of the above
8. Best model was evaluated by
  - (a) Adjusted R square
  - (b) Model Cp, AIC, BIC
  - (c) The standard error of regression mean
  - (d) All of the above
9. Independent effect was evaluated either in linear or logistic by
  - (a) t-test
  - (b) Wald chi-square test
  - (c) Z test
  - (d) All of the above
10. Interaction term in regression model allows to
  - (a) Independent effect of two variables
  - (b) Assess whether there is an interaction between two variables
  - (c) Unadjusted effect of two variables
  - (d) All of the above

### Answer Keys

1. (d)
2. (a)
3. (d)
4. (b)
5. (d)
6. (d)
7. (d)

8. (d)
9. (d)
10. (b)

## References

1. Frost J. Choosing the correct type of regression analysis. Available at: <https://statisticsbyjim.com/regression/choosing-regression-analysis/>
2. Sarangam A. Different types of regression analysis – a basic guide. Available at: <https://www.jigsawacademy.com/blogs/data-science/types-of-regression-analysis/>
3. Bhuiyan AR, Srinivasan SR, Chen W, Azevedo MJ, Berenson GS. Influence of low birth weight on C-reactive protein in asymptomatic younger adults: the Bogalusa Heart Study. BMC Res Notes. 2011;4(1):1–5.
4. Plichta SB, Kelvin E. Munro's statistical methods for health care research. 6th ed. Philadelphia: Lippincott Williams & Wilkins; 2013.
5. Glantz SA, Slinker BK, Neilands TB. Primer of applied regression and analysis of variance. 3rd ed. New York: McGraw-Hill, Inc.; 2001.
6. Elliott AC, Woodward WA. SAS essentials: mastering SAS for data analytics. 3rd ed. McGraw-Hill Education; 2016, Brooks/Cole, Cengage Learning.
7. Fox J. Regression diagnostics: an introduction. 2nd ed. Canada: McMaster University. Sage Publishing; 2019.
8. Zhou U. STAT 540: data analysis and regression. Colorado State University; 2015.
9. Emmert-Streib F, Dehmer M. Evaluation of regression models: model assessment, model selection and generalization error. Mach Learn Knowl Extr. 2019;1(1):521–51. <https://doi.org/10.3390/make1010032>.
10. Cohen R. Introducing the GLMSELECT PROCEDURE for model selection. Available at: <https://facweb.cdm.depaul.edu/sjost/csc423/documents/glmselect-summary.pdf>
11. Pagano M, Gauvreau K, Mattie H. Principles of Biostatistics. 3rd ed. Chapman & Hall; 2022, Belmont, CA 94002-3098.
12. Corraini P, Olsen M, Pedersen L, Dekkers OM, Vandenbroucke JP. Effect modification, interaction, and mediation: an overview of theoretical insights for clinical investigators. Clin Epidemiol. 2017;9:331–8. <https://doi.org/10.2147/CLEP.S129728>.
13. Bhalla D. Logistic regression analysis with SAS. Available at: <https://www.listendata.com/2013/04/logistic-regression-analysis-with-sas.html>
14. Zhang X. Maximum likelihood estimation (MLE) and the Fisher information. Available at: <https://towardsdatascience.com/maximum-likelihood-estimation-mle-and-the-fisher-information-1dd53faa369>
15. Rawlings JO, Pantula SG, Dickey DA. Applied regression analysis: a research tool (springer texts in statistics). 2nd ed. Dordrecht: Springer; 2001.
16. Rosner B. Study guide of fundamentals of biostatistics. Available at: <https://old.amu.ac.in/emp/studym/100018285.pdf>
17. Rodriguez G. Generalized regression models. Available at: <https://grodri.github.io/glms/notes/c3s8>
18. SAS Help Center. The LOGSELECT Procedure. Available at: [https://documentation.sas.com/doc/en/pgmsascdc/9.4\\_3.5/casstat/casstat\\_logselect\\_details17.htm](https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.5/casstat/casstat_logselect_details17.htm)
19. Elliott AC, Woodward WA. SAS essentials: mastering SAS for data analytics. 2nd ed. Newark: Wiley; 2015.
20. Hosmer DW Jr, Lemeshow SA, Sturdivant RX. Applied logistic regression. 3rd ed. Hoboken: Wiley; 2013.
21. Williams R. Ordinal independent variables. University of Notre Dame Available at: <https://www3.nd.edu/~rwilliam/stats3/OrdinalIndependent.pdf>

# Chapter 23

## Epidemiology of the COVID-19 Pandemic: An Update



Amal K. Mitra

### Learning Objectives

After completing this chapter, you will be able to:

- Describe the epidemiology of COVID-19, including infectiousness, pathogenicity, virulence, and risk factors
- Document complications of the disease, including long-COVID symptoms.
- Understand the impact of COVID-19 on mental health
- Assess premature deaths due to COVID-19 using potential years of life lost.
- Describe treatment protocols for COVID-19
- Suggest vaccine effectiveness

## 1 Introduction

In December 2019, a novel coronavirus designated SARS-CoV-2 emerged in humans in the city of Wuhan, China, and caused an outbreak of an unusual number of cases of viral pneumonia. Also known as coronavirus disease 2019 (COVID-19), the disease, being highly transmissible, has spread quickly as a pandemic all over the world. In the past, several outbreaks of the virus have been traced, originating in birds, pigs, bats, and other animals that mutated to become infectious and dangerous to humans.

As of February 3, 2023, the pandemic due to the coronavirus COVID-19 is affecting 229 countries and territories, causing about 676.3 million cases and more than 6.8 million deaths worldwide. In the United States alone, the devastating pandemic has caused 104.5 million cases and nearly 1.14 million deaths. Of the top five countries in terms of morbidity and mortality, India stands second, followed by France, Germany, and Brazil. The top five countries comprise 39% of the total cases and 39.8% of total deaths due to COVID-19 worldwide [1].

---

A. K. Mitra (✉)

Department of Epidemiology and Biostatistics, Jackson State University, Jackson, MS, USA  
e-mail: [amal.k.mitra@jsums.edu](mailto:amal.k.mitra@jsums.edu)

## 2 Infectiousness, Pathogenicity, and Virulence

COVID-19 has overwhelmingly surpassed two previously known large outbreaks due to coronavirus, severe acute respiratory syndrome (SARS) and middle east respiratory syndrome (MERS) in terms of both infectiousness (ability to pass easily from one person to another) and the spatial range of epidemic areas [2]. However, COVID-19 appeared to be less virulent in terms of fatality rate (2.3%) than that of SARS (9.5%) and MERS (34.4%) [3].

Several mutations of the virus changed its nature in terms of infectiousness (transmissibility), pathogenicity (causing disease), and virulence (disease severity). Numerous variants of the virus are being tracked in the United States and globally during this pandemic. According to the Centers for Disease Control and Prevention (CDC), the currently prevailing omicron variant (lineages of BA.2, BA.4, and BA.5) spreads more easily than other variants, including the Delta variant [4]. This high level of infectiousness of omicron has been shown in person-to-person transmission, regardless of vaccination status or whether or not they are symptomatic. Even the omicron variant has the potential to produce disease (high pathogenicity) in people who have recovered from COVID-19. However, omicron is less virulent, meaning it causes less severe illness and death compared to the Delta variant.

Clinical outcomes among 222,688 cases with omicron variant infections and 23,305 cases with delta variant infections were compared in a longitudinal study in Southern California between December 2021 and January 2022 [5]. The omicron variants were BA.1 or its sublineages. Adjusted hazard ratios of progression to hospital admission, symptomatic hospital admission, intensive care unit admission, mechanical ventilation, and death were 0.59 (95% CI: 0.51–0.69), 0.59 (0.51–0.68), 0.50 (0.29–0.87), 0.36 (0.18–0.72), and 0.21 (0.10–0.44), respectively, for cases with omicron, compared with delta variant infections [5].

## 3 Risk Factors

COVID-19 can affect anyone, but elderly people and people with comorbidities are more likely to get severe disease. CDC guidelines suggest that if multiple factors below indicate high transmission risk, you should consider adding more preventive actions.

***Factors that increase the risk of transmission*** Some of the known factors that increase the risk of COVID-19 transmission are as follows:

- **Length of exposure:** Longer the exposure time, the higher the risk of transmission.
- **Cough or heavy breathing:** Activities like coughing, breathing, shouting, or singing loudly and heavily increase the risk of transmission from an infected person to others.

- **Having symptoms:** Being around people with symptomatic illnesses increases the risk of transmission.
- **Not using masks:** Both the infected person and the people around him/her should wear a high-quality mask (such as N95) or a respirator. It is considered one of the best preventive measures against COVID-19.
- **Poor ventilation and filtration:** Indoor facilities with improper ventilation and filtration of air increase the risk. Being outside would lower the risk more than being indoors.
- **Not maintaining social distancing:** Being closer to an infected person increases the risk of transmission. If you are not fully vaccinated, practice social distancing by putting space (at least 6 feet) between yourself and others. Social distancing may make some people feel socially or culturally isolated and possibly lead to loneliness, depression, and poor health. However, it is useful to maintain other nonphysical ways to connect with family and friends, like sending a letter, making phone calls or video calls, or using social media.
- **Not washing hands:** Improper or no handwashing increases the risk of many bacterial, viral, and parasitic diseases, including COVID-19. Good hand-washing practice, meaning washing both hands with soap and water for at least 20 seconds, can prevent most infections. Handwashing should be done before, during, and after each activity to stay healthy. If soap and water are not available, use an alcohol-based hand rub that contains at least 60% alcohol.

### Other Risk Factors

- **Age:** Older adults are more likely to get severe infections and suffer more deaths. People with severe infections may need hospitalization, intensive care, or a ventilator. Mortality rates increase with ages over 50. People over 85 and older are at the highest risk of serious symptoms.
- **Race:** African Americans are contracting SARS-CoV-2 at higher rates and are more likely to die. In Chicago, more than 50% of COVID-19 cases and nearly 70% of COVID-19 deaths involve black individuals, although blacks make up only 30% of the population. Moreover, these deaths are concentrated mostly in just five neighborhoods on the city's South Side. In Louisiana, 70.5% of deaths have occurred among black people, who represent 32.2% of the state's population. In Michigan, 33% of COVID-19 cases and 40% of deaths have occurred among black individuals, who represent 14% of the population [6]. Black people have the highest COVID-19 death rate in the nation (68 per 100,000), followed by American-Indians or Alaska Natives (34 per 100,000), Hispanic or Latino people (32 per 100,000), Asian people (29 per 100,000), and White people (27 per 100,000) [7].
- **Associated comorbidities:** Hypertension, diabetes, obesity, asthma, concomitant cardiovascular diseases (including coronary artery disease and heart failure), and myocardial injury are important risk factors associated with worse outcomes [6].

## 4 Complications of COVID-19

In a recent study in Oman [8], the disease pathogenesis and mortality were identified in a cohort of 1462 confirmed cases (aged  $55 \pm 17$  years). Patients infected with the alpha COVID-19 variant type were more likely to have acute respiratory distress syndrome (ARDS) ( $p < 0.001$ ), stay longer in the hospital ( $p < 0.001$ ), and get admitted to the intensive care unit (ICU) ( $p < 0.001$ ). At the same time, those who had the omicron COVID-19 type were more likely to have a renal impairment ( $p < 0.001$ ) and less likely to be associated with noninvasive ventilation ( $p = 0.001$ ) compared with other COVID-19 variant types.

All-cause mortality was higher among patients having the delta (adjusted odds ratio [aOR], 1.8; 95% confidence interval (CI): 1.22–2.66;  $p = 0.003$ ) and omicron (aOR, 1.88; 95% CI: 1.09–3.22;  $p = 0.022$ ) COVID-19 variant types when compared to the initial COVID-19 (or alpha) variant. Old age (aOR, 1.05; 95% CI: 1.04–1.06;  $p < 0.001$ ), the presence of respiratory disease (aOR, 1.58; 95% CI: 1.02–2.44;  $p = 0.04$ ), ICU admission (aOR, 3.41; 95% CI: 2.16–5.39;  $p < 0.001$ ), lower estimated glomerular filtration rate (eGFR, a measure of kidney function) (aOR, 1.61; 95% CI: 1.17–2.23;  $p = 0.004$ ), and acute respiratory distress syndrome (ARDS, having severe breathing difficulties) (aOR, 5.75; 95% CI: 3.69–8.98;  $p < 0.001$ ) were also associated with higher mortality, while noninvasive ventilation requirements were associated with lower odds of death (aOR, 0.65; 95% CI: 0.46–0.91;  $p = 0.012$ ) [8].

### 4.1 COVID-19 Impact on Mental Health

In a recent publication, the CDC shed new light on how difficult the COVID-19 pandemic has been for high school students [9]. The data came from the CDC's first nationally representative survey of high school students during the pandemic. Fifty-five percent of high school students reported that they had experienced emotional abuse by adults at home, and 11% mentioned that they were physically abused by a parent or other adult in the family. About 29% reported that a parent or other adults in the family lost jobs during the pandemic. Lesbians, gays, bisexual youths, and females had experienced a greater level of poor mental health, emotional abuse, and attempted suicide compared with their counterparts.

A systematic review of 16 quantitative studies conducted in 2019–2021 with records of 40,076 individuals by Jones et al. [10] showed that adolescents of varying backgrounds experienced higher rates of anxiety, depression, and stress due to the pandemic. These studies were from Canada ( $n = 2$ ), China ( $n = 7$ ), Denmark ( $n = 1$ ), Germany ( $n = 1$ ), Japan ( $n = 1$ ), the Philippines ( $n = 1$ ), the United Kingdom ( $n = 1$ ), and the United States ( $n = 2$ ). Adolescents also had a higher frequency of using alcohol and cannabis during the COVID-19 pandemic. Stressful life events (such as deaths in the family and family members suffering from COVID-19), extended

home confinement, worry, and overuse of the internet and social media were possible factors in the poor mental health of adolescents during this pandemic.

In a cross-sectional survey conducted among 1002 adults (mean age =  $34.7 \pm 13.9$ ) infected with COVID-19 in Bangladesh, 48% were categorized as having moderate to severe depression. Based on multivariate regression analysis, depression during COVID-19 was positively associated with lower family income, poor health status, sleep disturbance, lack of physical activity, hypertension, asthma/respiratory problems, fear of COVID-19 re-infection, and persistent COVID-19 symptoms [11].

In a study in China [12], 359 children and 3254 adolescents were surveyed using an online questionnaire. About 91% of respondents clearly reported concerns about the epidemic; 805 (22%) respondents were suffering from depressive symptoms. The significant predictors of depression ( $n = 3613$ ) were smartphone addiction (Odds ratio (OR) = 1.4, 95% CI = 1.1–1.2), Internet addiction (OR = 1.8, 95% CI = 1.2–2.8), residents of Hubei province (the hardest-hit province of the epidemic) (OR = 3.1, 95% CI = 1.3–7.7), and family members or friends infected with COVID-19 (OR = 3.7, 95% CI = 1.0–13.8). The significant predictors of anxiety ( $n = 3613$ ) included: female ( $p = 0.002$ ), family member of a friend infected with COVID-19 ( $p < 0.001$ ), occupation of the mother involved in the epidemic ( $p = 0.007$ ), and poor coping skills ( $p = 0.002$ ).

## 4.2 Long-COVID or Post-COVID Conditions

About 10–20% of people after a COVID-19 infection develop mild and long-term effects after they are recovered from the initial infection. This is called a post-COVID-19 condition (PCC) or long COVID illness [13]. Symptoms of long COVID (or post-COVID condition) are variable and last for weeks, months, or even years after the initial illness [14]. Long-COVID is more common in people who have had severe COVID illnesses but anyone who gets the infection may be affected [13, 14]. Long COVID affects multiple organs, including the lung, heart, kidney, skin, and brain. The symptoms range from tiredness or fatigue, fever, and cough to disabilities. Long-term damage to the cardiovascular system, the brain, and the nervous system is getting more attention from long COVID illnesses. People with long COVID have an increased risk of heart disease, diabetes, blood clotting disorders, and neurological conditions compared with those without COVID-19 disease. Patients with long COVID experience the following symptoms:

- Tiredness or fatigue
- Postexertional malaise
- Fever
- Difficulty in breathing or shortness of breath
- Cough
- Chest pain
- Palpitations

- “Brain fog”, difficulty in thinking or concentrating
- Headache
- Sleep disturbances
- Lightheadedness
- Pins-and-needles feelings
- Change in smell and taste
- Depression or anxiety
- Diarrhea
- Stomach pain
- Joint or muscle pain
- Rash
- Changes in the menstrual cycle

A report from Italy found that 87% of people who recovered and were discharged from hospitals showed persistence of at least one symptom even after 60 days [15]. Of these, 32% had one or two symptoms, whereas 55% had three or more symptoms. The most commonly reported problems were fatigue (53.1%), worsened quality of life (44.1%), dyspnea (43.4%), joint pain (27.3%), and chest pain (21.7%).

Data from the United Kingdom Office for National Statistics (ONS) suggests that 11.7% of patients with long COVID illnesses report symptoms lasting for >12 weeks, with 12.3% of secondary school students reporting symptoms for more than 4 weeks. In the United Kingdom, the Real-time Assessment of Community Transmission (REACT-2) survey estimates that more than two million people have suffered from long COVID symptoms. Of them, about 750,000 had some limitations in day-to-day activities, and about 384,000 had symptoms lasting for more than 1 year [16].

Evidence supports the involvement of multiple organ systems, with fibrosis and inflammation in the lung, heart, kidneys, central nervous system, liver, adrenal glands, bone marrow, lymph nodes, and gastrointestinal tract. The COVID virus infection has also been associated with serious thrombotic complications, including strokes, pulmonary embolism, and cardiac injury [16].

### **4.3 “Brain Fog” After COVID-19 Infection**

Patients after recovery from infection with COVID-19 have reported transient or even lasting cognitive dysfunction, having symptoms of deficits in attention, executive functioning, language, processing speed, and memory, collectively referred to as “brain fog” [17]. In addition, the patients have an increased incidence of anxiety, depression, sleep disorders, and fatigue. An animal study, using a mouse model, explored how mild respiratory infections caused by SARS-CoV-2 could lead to neuroinflammation and subsequent brain damage [17].

Management of people with long COVID requires a multidisciplinary approach, including medications to relieve common symptoms such as cough, pain, or

headache; treatment for underlying problems; physiotherapy and relaxation techniques to help with fatigue and sleep problems; occupational therapy; and psychological support [13, 14].

## 5 Changing Nature of COVID-19 Due to Mutations and Multiple Variants

A virus sometimes replicates or makes copies of it, making some changes to the original virus, which is called “mutation.” A virus after one or several new mutations is referred to as a “variant” of the original virus [18]. In other words, a variant is a viral genome (or genetic code) that contains one or more mutations. SARS-CoV-2, the virus that causes COVID-19, constantly changes but tends to change more slowly than other viruses such as HIV or influenza viruses. The World Health Organization (WHO) used simple, easy-to-call labels for coronavirus variants using the letters of the Greek alphabet, starting with the alpha variant, which emerged in 2020. Since the beginning of the pandemic, the virus has mutated into a number of prominent variants, including alpha, beta, gamma, delta, and omicron.

Most changes after mutation have little to no impact on the virus’ properties. However, depending on where the changes are located in the virus’s genetic material, they may affect the virus’s properties, including transmissibility (the ability for transmission) or severity. For example, the COVID-19 virus, after several mutations, may become more virulent [18]. A new variant of a disease is called a variant of interest if the new variant is suspected to cause significant changes, such as affecting many people or spreading to many countries. A variant is called a variant of concern when it spreads more easily, causes more severe disease, appears with new clinical presentations, or the routine control measures against the disease are becoming ineffective. The CDC is continuously monitoring to identify if there is any new variant of COVID-19, and if it becomes a variant of concern in terms of high transmissibility, more virulence (or the ability to produce more severe disease or more deaths), and more resistance to public health measures, including vaccine effectiveness and disease prevention compared with the original strain of the virus [19].

Going back to the evolution of the variants, Delta (B.1.617.2) was first identified in India in late 2020. Soon it appeared to be the predominant variant of the coronavirus and became the variant of concern throughout the world. Delta caused more than twice as many infections as Alpha variants. In the United States, in June 2021, after a steady decline in COVID-19 cases and hospitalizations, the arrival of Delta caused a rapid increase in the trend of COVID cases [19]. Delta caused more severe disease than other variants in people who were not vaccinated. All the common vaccines in the United States (mRNA vaccines, such as Moderna and Pfizer/BioNTech; Johnson and Johnson vaccine) were considered highly effective against severe illness, hospitalizations, and death from delta variants.

It was November 26, 2021, when WHO declared that the world was facing a new variant of COVID-19 called omicron. The original omicron strain (BA.1) was first identified in Botswana and South Africa. By January 2023, a new subvariant of omicron, called XBB.1.5, proved to be highly transmissible and became the predominant strain of COVID-19 in the US. Omicron's subvariants are considered to be especially efficient in spreading the disease [19]. Data has suggested that the original omicron strain was less severe, in general, than previous variants. However, surges in cases due to the new omicron variant can significantly increase the number of hospitalizations and deaths.

The question is, can the new variant be protected by the currently available vaccines? In 2022, Pfizer-BioNTech and Moderna bivalent booster shots were approved for everyone 6 months of age and older. In addition to the original alpha and delta strains of COVID-19, these boosters are designed to protect against disease caused by the omicron subvariants BA.4 and BA.5. However, the BQ and XBB subvariants posed new challenges to the vaccine's effectiveness as well as some COVID-19 treatments [19]. Persons who received either one or two monovalent COVID-19 vaccine boosters had much lower neutralization activity against omicron subvariants (especially against BA.2.75.2, BQ.1.1, and XBB) [20].

## 6 Assessing the Disease Burden Due to COVID-19

The common metrics used for quantifying the overall disease burden and prioritizing causes of death are disability-adjusted life years (DALYs), potential years of life lost (PYLL), quality-adjusted life year (QALY), and disability-adjusted life expectancy (DALE), among others [21–24]. PYLL has been used previously in calculating deaths and comparing the health system performance of countries in cases of major killer diseases such as cancer, cerebrovascular disease, ischemic heart disease, accidents, and a few infectious diseases [25–27], but not in cases of pandemics such as COVID-19. One issue with using PYLL for assessing the disease burden of COVID-19 was that the existing formula of PYLL gives greater weight to deaths at a younger age, whereas the COVID-19 pandemic disproportionately affected the elderly population. In this context, Mitra and colleagues (2020) suggested a revised PYLL formula, which has been used successfully in several country situations for quantifying the disease burden in terms of premature deaths due to COVID-19 [28].

### 6.1 Use of PYLL: A Case Study

The following data are needed for calculating the premature human losses due to COVID-19 in a country by using the PYLL formula:

- Age-specific number of deaths (not death rates)

- Life expectancy at birth
- Total population of the country (for calculating rates)

PYLL for a disease can be compared among multiple countries, and a standardized rate of PYLL can be calculated using a standard population for such comparisons. In the following exercise, the CDC data for age-specific deaths of COVID-19 in the United States as of December 28, 2022, were used [29]. For the proposed revised formula, the current life expectancy at birth in 2022 (80 years) and a total population of 335,915,434 were used [30].

### Data to Be Used

$$\text{PYLL (original formula)} = \sum_{i=1}^{69} a_i d_i = \sum_{i=1}^{69} (70 - i - 0.5) * d_i, \text{ developed by Romeder}$$

and McWhinnie (1977) [27]

PYLL (revised formula) =  $\sum (80 - i - 0.5) * d_i$  (the life expectancy has been revised).

In the above formula,  $a_i$  = remaining years of life until the age limit (in the original formula, the age limit was 70 years; in the revised formula, the age limit is 80 years);  $d_i$  = number of observed deaths in each class interval; and  $i$  = the midpoint of the class interval of each age group, and 0.5 is a constant.

First, the remaining years of life were calculated for each age group by subtracting the midpoint of the class interval and the constant number of 0.5 from the upper age limit.

For example, in Table 23.1, the midpoint of the class interval of 1–17 years is 8.5. Taking age 80 as the upper limit, the remaining years of life are  $80 - 8.5 - 0.5 = 71$ .

Table 23.1 shows that as of December 28, 2022, a total of 9,807,712 person-years have been lost due to premature deaths due to COVID-19 in the United States. This translates to 2934 person-years lost per 100,000 people in the country, using the US total population of 335,915,434.

## 7 Treatment Protocol for COVID-19

There are several standard guidelines suggested by the National Institute of Health (NIH), CDC, and Infectious Diseases Society of America (IDSA) for the treatment and management of COVID-19 [31–33]. Table 23.2 provides the NIH-recommended therapies for nonhospitalized adults with COVID-19.

The NIH guidelines for the treatment of adults hospitalized for COVID-19 are provided in Table 23.3.

**Table 23.1** Potential years of life lost (PYLL) due to COVID-19 in the United States (data as of December 28, 2022)

Age (years)	Midpoint of class interval ( $i$ )	All COVID-19 deaths ( $d_i$ )	Remaining years ( $a_i$ ) ( $80 - i - 0.5$ )	PYLL at age limit 80 ( $a_i * d_i$ )
0–17	8.5	1407	71	99,897
18–29	23.5	6737	56	377,272
30–39	34.5	19,240	45	865,800
40–49	44.5	45,041	35	1,576,435
50–64	57.0	197,015	22.5	4,432,838
65–74	69.5	245,547	9	2,455,470
75–84	79.5	281,640	0	0
≥85	89.5	287,931	0	0
Total		1,084,558		
Total PYLL				9,807,712
PYLL/100,000				2934.39

**Table 23.2** Therapeutic management of nonhospitalized adults with COVID-19

Patient disposition	Panel's recommendations
All patients	Offer symptom management. The panel recommends against the use of dexamethasone or other systematic corticosteroids in the absence of another indication.
Patients who are at high risk of progression to severe COVID-19	<p>Preferred therapies (listed in order of preference):            Ritonavir-boosted nirmatrelvir (Paxlovid)            Remdesivir</p> <p>Alternative therapy (when the preferred therapies are not available):            Molnupiravir, but recommends against using it for pregnant women.</p>

## 8 Vaccines for COVID-19

Vaccines help prevent severe illness, hospitalization, and death. In a nationwide comparative study of protection conferred by natural immunity through previous infection and vaccination against symptomatic omicron infection, vaccination showed enhanced protection among persons who had a previous infection. Immunity resulting from previous infection and recent booster vaccination conferred the strongest protection, compared to immunity from previous infection alone [34]. Viral load is a prominent factor affecting infectivity and vaccine effectiveness. A comparative study of the viral load of SARS-CoV-2 delta and omicron variants following vaccination and previous infection was conducted among more than 460,000 individuals. While recent vaccination reduces the omicron viral load, the effect wanes rapidly. Scientists observed vaccine effectiveness in decreasing morbidity and mortality but found less effectiveness of the vaccine on the transmissibility of omicron and its rapid waning compared to the delta variant [35]. Furthermore,

**Table 23.3** Therapeutic management of adults hospitalized for COVID-19

Patient disposition	Clinical scenario	Recommendation	Anticoagulant therapy
<b>Hospitalized for reasons other than COVID-19</b>	Patients with mild to moderate COVID-19 who are at high risk of progressing to severe COVID-19	See Table 23.1: Therapeutic Management of Nonhospitalized Adults with COVID-19 (provided earlier).	For patients without an indication for therapeutic anticoagulation: Prophylactic dose of heparin, unless contraindicated; for pregnant patients
Hospitalized but does not require oxygen supplementation	All patients	The panel <b>recommends against</b> the use of <b>dexamethasone or other systemic corticosteroids</b> for the treatment of COVID-19	
	Patients who are at high risk of progressing to severe COVID-19	<b>Remdesivir</b>	
Hospitalized and requires conventional oxygen	Patients who require minimal conventional oxygen	<b>Remdesivir</b>	
	Most patients	Use <b>dexamethasone plus remdesivir</b> . If remdesivir cannot be obtained, use <b>dexamethasone</b>	
	Patients who are receiving dexamethasone and who have rapidly increasing oxygen needs and systemic inflammation	Add <b>oral baricitinib</b> or <b>IV tocilizumab</b> to 1 of the options above	

(continued)

**Table 23.3** (continued)

Patient disposition	Clinical scenario	Recommendation	Anticoagulant therapy
Hospitalized and requires high-flow nasal-cannula oxygen (HFNC) or noninvasive ventilation	Most patients	Promptly start 1 of the following, if not already initiated: Dexamethasone plus oral baricitinib; dexamethasone plus IV tocilizumab. If baricitinib, tofacitinib, tocilizumab, or sarilumab cannot be obtained: Dexamethasone. Add remdesivir to 1 of the options above in certain patients	For patients without an indication for therapeutic anticoagulation: Prophylactic dose of heparin, unless contraindicated; for pregnant patients. For patients who are started on a therapeutic dose of heparin in a non-ICU setting and then transferred to the ICU, the panel recommends switching to a prophylactic dose of heparin, unless there is another indication for therapeutic anticoagulation
Hospitalized and requires mechanical ventilation or extracorporeal membrane oxygen	Most patients	Promptly start 1 of the following, if not already initiated: Dexamethasone plus oral baricitinib; dexamethasone plus IV tocilizumab. If baricitinib, tofacitinib, tocilizumab, or sarilumab cannot be obtained: dexamethasone	

many commercially available monoclonal antibody preparations were ineffective against the omicron variant (BA.1 strain). This observation suggested a search for new vaccines and booster doses of the vaccine. The U.S. Food and Drug Administration approved an updated COVID-19 vaccine, called bivalent vaccine, in mid-2022 [36]. The bivalent vaccine includes a component of the original virus strain and a component of the omicron variant. A subsequent CDC report [37] showed that bivalent boosters provided significant additional protection against symptomatic SARS-CoV-2 infection in patients who had previously received two, three, or four doses of monovalent vaccine. Due to the waning immunity of monovalent doses, the benefit of the bivalent booster is supposed to increase with time. The emergence of the omicron variant further emphasizes the importance of vaccination and boosters.

## 9 Further Practice

1. Define the following terms:
  - (a) Infectivity
  - (b) Pathogenicity
  - (c) Virulence
2. Choose the correct answer:
  - (a) COVID-19 virus is less virulent and less infectious than SARS and MERS viruses
  - (b) COVID-19 virus is less virulent but more infectious than SARS and MERS viruses
  - (c) COVID-19 virus is more virulent and more infectious than SARS and MERS viruses
3. Omicron variant is less virulent compared to the delta variant.      True/False
4. Long COVID shows the following symptoms:
  - (a) Diarrhea
  - (b) Difficulties of breathing or shortness of breath
  - (c) Brain fog
  - (d) Anxiety or depression
  - (e) All of the following
5. Common metrics used for the overall disease burden and prioritizing causes of deaths include:
  - (a) Disability-adjusted life years (DALYs)
  - (b) Potential years of life lost (PYLL)
  - (c) Quality-adjusted life-year (QALY)
  - (d) Disability-adjusted life expectancy (DALE)
  - (e) All of the above
6. Steroids, such as dexamethasone is recommended for the following conditions with COVID-19:
  - (a) Nonhospitalized patients
  - (b) Hospitalized patient but does not require oxygen supplementation
  - (c) Hospitalized patient and requires conventional oxygen
  - (d) All patients
7. Ramdisivir, an anti-viral medicine is recommended for the following patients with COVID-19, EXCEPT:
  - (a) Patients who are at high risk of progression to severe COVID-19
  - (b) Hospitalized patient who does not require oxygen supplementation but at high risk of progressing to severe COVID-19

- (c) Hospitalized patient and requires conventional oxygen
  - (d) Nonhospitalized patients
8. COVID-19 vaccines cannot prevent severe illness, hospitalization, and death      True/False
9. What is the difference between a monovalent and bivalent COVID-19 vaccine?  
Given an example for each.
10. Why do you need a booster? Please explain it in your own words.

### **Answers Keys**

1. See Sect. 2: Infectiousness, Pathogenicity, and Virulence
2. (b)
3. True
4. (e)
5. (e)
6. (c)
7. (d)
8. False
9. See Sect. 8: Vaccines for COVID-19
10. See Sect. 8: Vaccines for COVID-19

## **References**

1. Worldometer. Coronavirus. Available at: <https://www.worldometers.info/coronavirus/>. Accessed 5 Feb 2023.
2. Hu B, Guo H, Zhou P, et al. Characteristics of SARS-CoV-2 and COVID-19. *Nat Rev Microbiol*. 2021;19:141–54. <https://doi.org/10.1038/s41579-020-00459-7>.
3. Petrosillo N, Viceconte G, Ergonul O, Ippolito G, Petersen E. COVID-19, SARS and MERS: are they closely related? *Clin Microbiol Infect*. 2020;26(6):729–34. <https://doi.org/10.1016/j.cmi.2020.03.026>.
4. Centers for Disease Control and Prevention. COVID-19: variants of the virus. Available at: <https://www.cdc.gov/coronavirus/2019-ncov/variants/index.html>. Accessed 5 Jan 2023.
5. Lewnard JA, Hong VX, Patel MM, Kahn R, Lipsitch M, Tartof SY. Clinical outcomes associated with SARS-CoV-2 Omicron (B.1.1.529) variant and BA.1/BA.1.1 or BA.2 subvariant infection in Southern California. *Nat Med*. 2022;28(9):1933–43. <https://doi.org/10.1038/s41591-022-01887-z>.
6. Yancy CW. COVID-19 and African Americans. *JAMA*. 2020;323(19):1891–2. <https://doi.org/10.1001/jama.2020.6548>.
7. Zhang L, Mcleod ST, Vargas R, Liu X, Young DK, Dobbs TE. Subgroup comparison of COVID-19 case and mortality with associated factors in Mississippi: findings from analysis of the first four months of public data. *J Biomed Res*. 2020;34(6):446–57. <https://doi.org/10.7555/JBR.34.20200135>.
8. Khamis F, Al Awaidy S, Ba’Omar M, et al. The impact of demographic, clinical characteristics and the various COVID-19 variant types on all-cause mortality: a case-series retrospective study. *Diseases*. 2022;10:100. <https://doi.org/10.3390/diseases10040100>.
9. Centers for Disease Control and Prevention. New CDC data illuminate youth mental health threats during the COVID-19 pandemic. 2022. Available at: <https://www.cdc.gov/media-releases/2022/p0331-youth-mental-health-covid-19.html#print>. Accessed 4 Jan 2023.

10. Jones J, Mitra AK, Bhuiyan AR. Impact of COVID-19 on mental health of adolescents: a systematic review. *Int J Environ Res Public Health.* 2021;18:2470. <https://doi.org/10.3399/ijerph18052470>.
11. Islam MS, Ferdous MZ, Islam US, Mosaddek ASM, Potenza MN, Pardhan S. Treatment, persistent symptoms, and depression in people infected with COVID-19 in Bangladesh. *Int J Environ Res Public Health.* 2021;18:1453. <https://doi.org/10.3390/ijerph18041453>.
12. Duan L, Shao X, Wang Y, Huang Y, Miao J, Yang X, Zhu G. An investigation of mental health status of children and adolescents in China during the outbreak of COVID-19. *J Affect Disord.* 2020;275:112–8. <https://doi.org/10.1016/j.jad.2020.06.029>.
13. Raveendran AV, Jayadevan R, Sashidharan B. Long COVID: an overview. *Diabetes Metab Syndr Clin Res Rev.* 2021;15(3):869–75. <https://doi.org/10.1016/j.dsx.2021.04.007>.
14. CDC. Long COVID or post-COVID conditions. Available at: <https://www.cdc.gov/coronavirus/2019-ncov/long-term-effects/index.html#:~:text=Some%20people%2C%20especially%20those%20who,kidney%2C%20skin%2C%20and%20brain>. Accessed on 4 Jan 2023.
15. Carfi A, Bernabei R, Landi F, et al. Persistent symptoms in patients after acute COVID-19. *JAMA.* 2020;324(6):603–5. <https://doi.org/10.1001/jama.2020.12603>.
16. Desforges M, Gurdasani D, Hamdy A, Leonardi AJ. Uncertainty around the long-term implications of COVID-19. *Pathogens.* 2021;10:1267. <https://doi.org/10.3390/pathogens10101267>.
17. Venkataramani V, Winkler F. Cognitive deficits in long Covid-19. *N Engl J Med.* 2022;382:1813–5. <https://doi.org/10.1056/NEJMcibr2210069>.
18. World Health Organization. Coronavirus disease (COVID-19): virus evolution. Available at: <https://www.who.int/news-room/questions-and-answers/item/sars-cov-2-evolution>. Accessed 15 Feb 2023.
19. Katella K. Omicron, Delta, Alpha, and more: what to know about the Coronavirus variants. Yale Medicine; 2023. Available at: <https://www.yalemedicine.org/news/covid-19-variants-of-concern-omicron>. Accessed 15 Feb 2023.
20. Davis-Gardner ME, Lai L, Wali B, Samaha H, Solis D, Lee M, et al. Neutralization against BA.2.75.2, BQ.1.1, and XBB from mRNA Bivalent Booster. *N Engl J Med.* 2023;388(2):183–5. <https://doi.org/10.1056/NEJMcc2214293>.
21. World Health Organization. Disability-Adjusted Life Year (DALY). Available at: <https://www.who.int/data/gho/indicator-metadata-registry/imr-details/15#:~:text=Definition%3A,One%20DALY%20represents%20the%20loss%20of%20the%20equivalent%20of%20one,health%20condition%20in%20a%20population>. Accessed 15 Feb 2023.
22. Grandjean P, Bellanger M. Calculation of the disease burden associated with environmental chemical exposures: application of toxicological information in health economic estimation. *Environ Health.* 2017;16:123.
23. Egunsola O, Raubenheimer J, Buckley N. Variability in the burden of disease estimates with or without age weighting and discounting: a methodological study. *BMJ Open.* 2019;9:e027825.
24. Vienonen MA, Jousilahti PJ, Mackiewicz K, Oganov RG, Pisaryk VM, Denissov GR, et al. Preventable premature deaths (PYLL) in northern dimension partnership countries 2003–13. *Eur J Pub Health.* 2019;29:626–30.
25. Canadian Institute for Health Information. Potential years of life lost: international comparisons [product release]. Accessed 14 Feb 2023.
26. Maximova K, Rozen S, Springett J, Stachenko S. The use of potential years of life lost for monitoring premature mortality from chronic diseases: Canadian perspectives. *Can J Public Health.* 2016;107:e202–4.
27. Romeder JM, McWhinnie JR. Potential years of life lost between ages 1 and 70: an indicator of premature mortality for health planning. *Int J Epidemiol.* 1977;6:143–51.
28. Mitra AK, Payton M, Kabir N, Whitehead A, Ragland KN, Brown A. Potential years of life lost due to COVID-19 in the United States, Italy, and Germany: an old formula with newer ideas. *Int J Environ Res Public Health.* 2020;17:4392.
29. Statista. <https://www.statista.com/statistics/1191568/reported-deaths-from-covid-by-age-us/>
30. US Census. Bureau. Available at: <https://www.census.gov/>. Accessed 28 Dec 2022.

31. National Institutes of Health. COVID-19 treatment guidelines. Available at: [https://www.covid19treatmentguidelines.nih.gov/management/clinical-management-of-adults/clinical-management-of-adults-summary/?utm\\_source=site&utm\\_medium=home&utm\\_campaign=highlights](https://www.covid19treatmentguidelines.nih.gov/management/clinical-management-of-adults/clinical-management-of-adults-summary/?utm_source=site&utm_medium=home&utm_campaign=highlights). Accessed 15 Feb 2023.
32. Centers for Disease Control and Prevention. COVID-19 treatment and management. Available at: <https://www.cdc.gov/coronavirus/2019-ncov/your-health/treatments-for-severe-illness.html>. Accessed 15 Feb 2023.
33. Infectious Diseases Society of America. IDSA guidelines on the treatment and management of patients with COVID-19. Available at: <https://www.idsociety.org/practice-guideline/covid-19-guideline-treatment-and-management/#OverviewofCOVID-19TreatmentGuidelinesSummaryTable>. Accessed 15 Feb 2023.
34. Altarawneh HN, Chemaitelly H, Ayoub HH, et al. Effects of previous infection and vaccination on symptomatic omicron infections. *N Engl J Med.* 2022;387:21–34. <https://doi.org/10.1056/NEJMoa2203965>.
35. Barouch DH. Covid-19 vaccines – immunity, variants, boosters. *N Engl J Med.* 2022;387:1011–22.
36. U.S. Food & Drug Administration. Covid-19 bivalent vaccine boosters. Available at: <https://www.fda.gov/emergency-preparedness-and-response/coronavirus-disease-2019-covid-19/covid-19-bivalent-vaccine-boosters>. Accessed 15 Feb 2023.
37. Link-Gelles R, Ciesla AA, Roper LE, et al. Early estimates of bivalent mRNA booster dose vaccine effectiveness in preventing symptomatic SARS-CoV-2 infection attributable to Omicron BA.5– and XBB/XBB.1.5–related sublineages among immunocompetent adults — increasing community access to testing program, United States, December 2022–January 2023. *MMWR Morb Mortal Wkly Rep.* 2023;72:119–24. <https://doi.org/10.15585/mmwr.mm7205e1>.

# Index

## A

- Age-specific death rate, 149–151  
Analogy, 163  
Analysis of variance (ANOVA), 346, 347, 356, 358–361, 364, 365, 367, 375, 379, 388  
Analytical studies, 2, 7, 10–15, 21, 23, 34

## B

- Bias, 27, 45, 66, 96, 121, 151, 169, 194, 225, 248, 258, 280, 304, 329, 390  
Binary logistic regression, 312, 314, 391–397  
Bogalusa Heart Study (BHS), 382

## C

- Case, 7, 20, 43, 68, 78, 99, 116, 131, 150, 157, 170, 183, 219, 236, 259, 294, 322, 344, 372, 382, 411  
Case-control study, 45  
Causal association, 155–165, 174  
Censored data, 235, 237–239, 246, 248  
Chi-square test, 246, 356, 360, 361, 398, 402, 408  
Classification, 20, 35, 82, 122, 222, 267, 297–299, 317, 322–324, 332–334, 336, 339–341  
Clinical trials, 20, 91–101, 103–105, 107, 108, 161–163, 276, 283–285, 290  
Clustering, 16, 19, 20, 192, 317, 325–329, 341  
Coherence, 162  
Cohort studies, 9, 10, 13–15, 26, 36, 37, 47, 48, 57–72, 156, 158, 160, 171, 193–194, 280, 281, 288, 291

- Collinearity, 383, 386–387, 396, 408  
Confidence, 47, 265, 282, 284–291, 373, 378  
Confidence intervals (CIs), 13, 43, 49–53, 105, 158, 159, 161, 180, 266, 358, 359, 373, 374, 378, 404, 412, 414, 415

- Confounder, 47, 100, 156, 157, 173, 175, 248  
Confounding, 11, 14, 23, 24, 37, 48, 49, 60, 71, 99, 100, 132, 155–157, 169, 171, 173–178, 180

- Consistency, 98, 160–161, 285  
Correlation, 11, 23, 47, 48, 50, 52, 225, 267, 309, 322, 357, 360, 361, 364, 365, 368, 376–379, 382, 386, 396, 398, 407

- COVID-19, 15, 36, 78, 86, 87, 125, 130, 152, 184, 199, 217, 239, 411–424  
Cross-sectional studies, 7, 10, 12, 15, 16, 20, 24–31, 34–39, 155, 156, 159, 160, 276, 281, 282, 288, 289  
Crude and age-adjusted rates, 77, 78

## D

- Data collection, 22, 25, 26, 34–36, 50, 62, 66, 77, 132, 135, 136, 140, 160, 187  
Deep learning (DL), 318, 328, 340  
Delta variant, 412, 417, 420, 423  
Demography, 206, 207, 241  
Descriptive statistics, 78, 347–349, 352, 353, 363, 366, 370, 379  
Descriptive studies, 7, 20, 23, 24  
Direct method, 132, 148, 151  
Disease, 1, 19, 43, 58, 77, 115, 129, 155, 171, 184, 207, 219, 235, 267, 286, 319, 345, 364, 411

Disease outbreak, 11, 130, 196, 199, 227  
Dose-response, 161

**E**

Ecological studies, 2, 10, 11, 15, 16, 22, 23, 78  
Effect modification, 176–180, 391  
Effect size, 263, 276–279, 290, 291, 374, 375  
Epidemic, 29, 130, 131, 164, 183–189,  
    191–193, 197–200, 412, 415  
Error rates, 47, 276, 277, 279  
Estimation, 130, 194, 204, 242, 261, 270,  
    281–290, 305, 381  
Evidence-based decision-making, 257–271  
Expected death rates, 148, 150, 157  
Experimentation, 162–163

**F**

Fertility, 78, 80, 203, 204, 206, 207, 211

**G**

Geographic factors, 197, 220–222  
Geographic mapping, 2, 6, 28, 80, 87, 184,  
    217, 228  
G\*Power, 276, 288–290

**H**

Handling missing data, 297  
Hazard function, 235, 239, 241, 242,  
    245–248, 252–254  
Hierarchy of studies, 217  
Hill's criteria, 155–163  
Hypothesis, 7, 9, 11, 12, 16, 20, 23, 30, 34, 44,  
    47, 65, 66, 93, 95–97, 102–105,  
    136, 160, 162, 163, 188, 190, 193,  
    194, 197, 219, 250, 259, 269,  
    276–281, 291, 364, 365, 371–379,  
    383, 388–389, 398, 400

**I**

Imputation, 294, 303–315  
Incidence, 3, 5–7, 9, 10, 13, 21, 22, 24, 36, 38,  
    48, 58, 64–69, 77, 78, 85–87, 95,  
    134, 136, 138, 158, 160, 161, 169,  
    185, 186, 219, 222, 231, 282, 416  
Indirect method, 132  
Inferential statistics, 364, 378  
Information bias, 46, 48, 64, 66–68, 121,  
    169, 171–173  
Informed consent, 105–109

Intention-to-treat analysis, 95, 103, 104

**J**

Jackson Heart Study (JHS), 363–365, 367, 368  
Kaplan–Meier, 242–247, 250–253

**L**

Likelihood ratio, 250, 397, 398, 408  
Log rank test, 242, 246–248, 251  
Long-Covid, 415, 416, 423

**M**

Machine learning (ML), 294, 317–328, 330,  
    340, 341  
Meta-analysis, 103, 105, 257–263,  
    265–271  
Migration, 81, 203, 204, 206–208, 211,  
    212, 221  
Missing at random (MAR), 297–299, 303,  
    309, 314  
Missing completely at random (MCAR),  
    297–305, 314  
Missing not at random (MNAR), 297, 299,  
    303, 314  
Missing value analysis, 68, 293–295, 299,  
    304, 309, 310, 314

Model fit assumption, 383–389, 395  
Models, 162, 204, 220, 241, 265, 271,  
    317–320, 322, 323, 329–331, 337,  
    339–341, 391

Morbidity, 6, 77, 78, 84–87, 116, 132, 134,  
    152, 236, 243, 246, 248–250, 252,  
    411, 420

Mortality, 3, 6, 7, 22, 23, 57, 58, 60, 66, 72,  
    78–88, 116–117, 122, 132, 134,  
    152, 162, 203, 204, 206, 207, 211,  
    221, 241, 264, 267–268, 411, 413,  
    414, 420  
Multiple linear regression, 357, 361, 381, 382,  
    387, 388

Multiple logistic regression, 357, 361,  
    391, 395–398

**O**

Observational study, 10, 43, 57, 61, 176,  
    222, 261  
Odds ratio (OR), 11–13, 16, 32, 33, 37, 38, 43,  
    47, 50–53, 158, 159, 169–171,  
    176–179, 194, 281, 286–287, 291,  
    404, 414, 415

Omicron variant, 412, 418, 420, 422, 423  
Outbreak investigation, 136, 140,  
183, 188–199

## P

Paired *t*-test, 355, 357  
Person–place–time model, 2  
Plausibility, 91, 162  
Population projection, 203–212  
Power, 46–48, 65, 68, 93, 95, 221, 276, 278,  
279, 284, 285, 289–291, 297,  
318, 353  
Predictive value, 119, 139, 400  
Prevalence, 3, 22, 43, 66, 78, 95, 134, 169,  
207, 219, 282, 364  
Prevention and control, 2, 7, 129  
Probabilistic projections, 206, 208  
Prospective study, 62, 197  
Public health, 3, 5, 7, 8, 14, 15, 20, 21, 28, 30,  
34, 36, 67, 71, 78–82, 84, 87, 88,  
94, 100, 115, 116, 129–144, 152,  
187, 189, 190, 196, 203, 204,  
207–208, 221, 257, 258, 260, 265,  
269, 270, 286, 381, 417

## R

Random allocation, 98, 175  
Randomized controlled trials (RCTs), 14, 15,  
28, 36, 63, 93, 96, 98, 172, 175, 281  
Real-life examples, 115, 165, 183,  
276, 294–296  
Regression, 13, 157, 174–176, 237, 267, 304,  
305, 307–310, 313, 315, 317–323,  
328, 330, 331, 340, 378, 381–408, 415  
Regression tests, 267  
Reliability, 117, 118, 140, 236  
Repeated analysis, 117, 160  
Retrospective study, 62  
Risk factors, 1, 3, 4, 7, 10–16, 25, 26, 28–30, 33,  
35, 36, 38, 45, 47, 49–51, 61, 63–66,  
129, 135, 136, 138, 155–157, 159,  
160, 163–165, 171, 173, 177, 194,  
248, 294–296, 311, 364, 412–413

## S

SARS-CoV-2, 411, 413, 416, 417, 420, 422  
SAS, 235, 243, 246, 249–254, 263, 271, 304,  
383, 386–387, 390–391,  
396–402, 404–407

Scikit-learn, 319–321, 323, 324, 326, 334  
Screening, 25, 49, 84, 115–126, 345  
Selection bias, 46, 48, 64, 66, 67, 98, 169–172,  
175, 194

Sensitivity, 26, 104, 117–119, 122, 125, 126,  
197, 260, 265, 271, 328, 399

Simple linear regression, 320, 382

Specificity, 96, 117–119, 122, 125, 126,  
137, 161–162

SPSS data analysis, 235, 263, 271, 304,  
343–361, 363–379

Standardization, 60, 130, 132, 147–153

Standardized mortality ratio (SMR),  
60, 67, 152

Statistical package for the social sciences  
(SPSS), 235, 246, 250–254, 263,  
271, 304, 343, 347–354, 357, 358,  
365–373, 375–377

Steps of epidemic investigation, vii, 183–199

Strength of association, 51, 157–159, 194, 377

Student *t*-test, 347, 371

Study design, 1, 7, 9–12, 16, 20–23, 25, 27,  
34, 36, 43, 44, 57, 58, 62, 63,  
93–96, 155, 158, 160, 169, 276,  
281, 288

Surveillance, 9, 21, 65, 129, 187

Surveillance systems, 29, 129, 130, 135–142,  
189, 196, 197

Survival analysis, 13, 235–251, 253

Systematic review, 257–263, 265,  
268–271, 414

## T

Temporality, 156, 159–160  
TensorFlow, 330, 332, 333, 335, 338  
Test of normality, 353–354, 383–385  
Therapeutic equipoise, 91, 94, 97, 104  
Time-to-event analysis, 235–237, 248, 251,  
262, 270  
Treatment, 4, 13, 14, 20–22, 28, 50, 84, 93–95,  
98, 99, 103–105, 108, 115–117,  
130, 161–163, 198, 220, 261, 270,  
276, 278, 283–285, 290, 322,  
344–346, 350, 351, 353,  
356–360, 417–421

## V

Vaccination, 14, 25, 51, 196, 412, 420, 422  
Validity, 16, 46, 98, 106, 117–118,  
170–172, 265